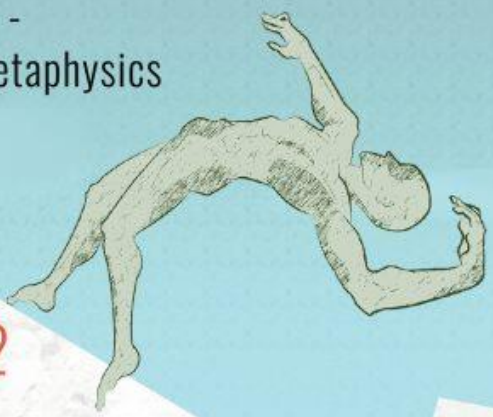
The Rational Human Condition

Deep Freedom and
Real Persons -
A Study in Metaphysics



VOLUME 2

World Philosophy

ROBERT HANNA

NOVA

WORLD PHILOSOPHY

THE RATIONAL HUMAN CONDITION

VOLUME 2

DEEP FREEDOM AND REAL PERSONS: A STUDY IN METAPHYSICS

WORLD PHILOSOPHY

Cover Art: “The Human Condition Rationalized,” by Otto Paans

The first four hard-copy volumes in THE RATIONAL HUMAN CONDITION series can be found on Nova’s website, **HERE**.

The first four e-books in THE RATIONAL HUMAN CONDITION series can be found on Nova’s website, **HERE**.

WORLD PHILOSOPHY

THE RATIONAL HUMAN CONDITION

VOLUME 2

DEEP FREEDOM AND REAL PERSONS: A STUDY IN METAPHYSICS

ROBERT HANNA

Library of Congress Cataloging-in-Publication Data

ISBN: 978-1-53614-519-9

Published by Nova Science Publishers, Inc. † New York

All five volumes of THE RATIONAL HUMAN CONDITION are dedicated
to the people I love—you know who you are.
But especially Martha and Beth.
And also to all those who helped me with its ideas and arguments.

CONTENTS

"The Human Condition," by Thomas Whitaker/Prison Arts Coalition.

PREFACE

Robert Hanna's THE RATIONAL HUMAN CONDITION is a five-volume book series, including:

- Volume 1. *Preface and General Introduction, Supplementary Essays, and General Bibliography*
- Volume 2. *Deep Freedom and Real Persons: A Study in Metaphysics*
- Volume 3. *Kantian Ethics and Human Existence: A Study in Moral Philosophy*
- Volume 4. *Kant, Agnosticism, and Anarchism: A Theological-Political Treatise*
- Volume 5. *Cognition, Content, and the A Priori: A Study in the Philosophy of Mind and Knowledge*

The fifth volume in the series, *Cognition, Content, and the A Priori*, was published by Oxford University Press in 2015. So, with the present publication of the first four volumes in the series by Nova Science in 2019, all five volumes of THE RATIONAL HUMAN CONDITION are now available in hard-copy and as e-books. All five books share a common aim, which is to work out a true general theory of human rationality in a thoroughly nonideal natural and social world. This philosophical enterprise is what Hanna calls rational anthropology. In the eleventh and most famous of his Theses on Feuerbach, Karl Marx wrote that "philosophers have only interpreted the world in different ways; the point is to change it." Hanna completely agrees with Marx that the ultimate aim of philosophy is to change the world, not merely interpret it. So, Marx and Hanna are both philosophical liberationists: that is, they both believe that philosophy should have radical political implications. But, beyond Marx, Hanna also thinks that the primary aim of philosophy (understood as rational anthropology) and its practices of synoptic reflection, writing, teaching, and public conversation is to change lives for the better—and ultimately, for the sake of the highest good. Then, and only then, can the human race act upon the world in the right way. The first four volumes of THE RATIONAL HUMAN CONDITION will therefore appeal not only to philosophers, but also to any other philosophically-minded person interested in the intellectual and practical adventure of synoptic, reflective thinking about the nature of our rational, but still ineluctably "human, all-too-human" lives.

A NOTE ON REFERENCES

Throughout the five-volume series THE RATIONAL HUMAN CONDITION, for convenience, I refer to Kant's works infratextually in parentheses. The references include both an abbreviation of the English title and the corresponding volume and page numbers in the standard "Akademie" edition of Kant's works: *Kants gesammelte Schriften*, edited by the Königlich Preussischen (now Deutschen) Akademie der Wissenschaften (Berlin: G. Reimer [now de Gruyter], 1902-). I generally follow the standard English translations, but have occasionally modified them where appropriate. For references to the first *Critique*, I follow the common practice of giving page numbers from the A (1781) and B (1787) German editions only. Here is a list of the relevant abbreviations and English translations:

BL "The Blomberg Logic." In *Immanuel Kant: Lectures on Logic*. Trans. J.M. Young. Cambridge: Cambridge Univ. Press, 1992. Pp. 5-246.

C *Immanuel Kant: Correspondence, 1759-99*. Trans. A. Zweig. Cambridge: Cambridge Univ. Press, 1999.

CF *Conflict of the Faculties*. Trans. M. Gregor. Lincoln, NE: Univ. of Nebraska Press, 1979.

CPJ *Critique of the Power of Judgment*. Trans. P. Guyer and E. Matthews. Cambridge: Cambridge Univ. Press, 2000.

CPR *Critique of Pure Reason*. Trans. P. Guyer and A. Wood. Cambridge: Cambridge Univ. Press, 1997.

CPrR *Critique of Practical Reason*. Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy*. Cambridge: Cambridge Univ. Press, 1996. Pp. 139-271.

DiS "Concerning the Ultimate Ground of the Differentiation of Directions in Space." Trans. D. Walford and R. Meerbote. In *Immanuel Kant: Theoretical Philosophy: 1755-1770*. Cambridge: Cambridge Univ. Press, 1992. Pp. 365-372.

DSS "Dreams of a Spirit-Seer Elucidated by Dreams of Metaphysics." Trans. D. Walford and R. Meerbote. In *Immanuel Kant: Theoretical Philosophy: 1755-1770*. Pp. 301-359.

EAT "The End of All Things." Trans. A. Wood and G. Di Giovanni. In *Immanuel Kant: Religion and Rational Theology*. Cambridge: Cambridge Univ. Press, 1996. Pp. 221-231.

GMM *Groundwork of the Metaphysics of Morals*. Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy*. Pp. 43-108.

ID "On the Form and Principles of the Sensible and Intelligible World (Inaugural Dissertation)." Trans. D. Walford and R. Meerbote. In *Immanuel Kant: Theoretical Philosophy: 1755-1770*. Pp. 373-416.

IUH "Idea for a Universal History with a Cosmopolitan Aim." Trans. A. Wood. In *Immanuel Kant: Anthropology, History, and Eduction*. Cambridge: Cambridge Univ. Press, 2007. Pp. 107-120.

JL "The Jäsche Logic." Trans. J.M. Young. In *Immanuel Kant: Lectures on Logic*. Pp. 519-640.

LE *Immanuel Kant: Lectures on Ethics*. Trans. P. Heath. Cambridge: Cambridge Univ. Press, 1997.

MFNS *Metaphysical Foundations of Natural Science*. Trans. M. Friedman. Cambridge: Cambridge Univ. Press, 2004.

MM *Metaphysics of Morals*. Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy*. Pp. 365-603.

OP *Immanuel Kant: Opus postumum*. Trans. E. Förster and M. Rosen. Cambridge: Cambridge Univ. Press, 1993.

OPA "The Only Possible Argument in Support of a Demonstration of the Existence of God." Trans. D. Walford and R. Meerbote. In *Immanuel Kant: Theoretical Philosophy: 1755-1770*. Pp. 107-201.

OT "What Does It Mean to Orient Oneself in Thinking?" Trans. A. Wood. In *Immanuel Kant: Religion and Rational Theology*. Pp. 7-18.

Prol *Prolegomena to Any Future Metaphysics*. Trans. G. Hatfield. Cambridge: Cambridge Univ. Press, 2004.

PP "Toward Perpetual Peace." Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy*. Pp. 317-351.

Rel *Religion within the Boundaries of Mere Reason*. Trans. A. Wood and G. Di Giovanni. In *Immanuel Kant: Religion and Rational Theology*. Pp. 57-215.

RTL "On a Supposed Right to Lie from Philanthropy." Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy*. Pp. 611-615.

VL "The Vienna Logic," Trans. J.M. Young. In *Immanuel Kant: Lectures on Logic*. Pp. 251-377.

WiE "An Answer to the Question: 'What is Enlightenment?'" Trans. M. Gregor. In *Immanuel Kant: Practical Philosophy*. Pp. 17-22.

Chapter 1

INTRODUCTION: FREEDOM, LIFE, AND PERSONS' LIVES

We … cognize practical freedom through experience as one of the natural causes, namely a causality of reason in the determination of the will, whereas transcendental freedom requires an independence of this reason itself (with regard to its causality in initiating a series of appearances) from all determining causes of the world, and to this extent seems to be contrary to the law of nature, thus to all possible experience, and so remains a problem. (*CPR* A803/B831)

Therein is contained the whole wisdom of life, but no one has ever rendered them as impressively—as if he were a god in the shape of a scarecrow who spoke to suffering humanity—as that great thinker and genuine philosopher of life who said to a man who had hurled his hat to the floor: Pick it up, and you will get a beating; leave it there and you will also get a beating; now you may choose. You have your great joy "comforting" people when they turn to you in crucial situations; you listen to their expositions and then say: Yes, now I see it all perfectly; there are two possible situations—one can do either this or that. My honest opinion and my friendly advice is this: Do it or do not do it—you will regret both….[Your] view of life is concentrated in one single sentence: I say simply Either/Or.[1]

What has to be accepted, the given, is—so one could say—forms of life.[2]

There is an absolute contradiction between the freedom we all presuppose in practice and the implications of ideas that are widely accepted as established scientific fact. Philosophy has no higher calling than to try to resolve this contradiction at the heart of contemporary culture.[3]

If one accepts classical physics, free will must apparently be explained as being *compatible* with determinism. The only alternative to compatibilism, if sense is to be made of free will, would be to postulate that the laws of physics do not have universal application and the human free will can cause things to happen contrary to those laws. It might be suggested that Kant found a third alternative, but if so it is one I am unable to understand.[4]

I am convinced that to solve the problems surrrounding animal agency simply it is not enough to cast off mistaken theories of causation…. It needs also to be shown that real, biological processes might enable us to sustain the idea that an *animal* may be truly in charge of what it does, so that its actions are more than merely the byproduct of its innards

> and parts. The task requires some reflection on the organizational principles of living creatures, for it is only through such reflection … that we can start to understand where the difference really lies between, on the one hand those things that are true agents, and, on the other, mere machines, entities that nothing will ever be up to, however impressive they may be…. I am exceedingly hopeful that the next few years will see the beginnings of a revolution in our conception of the human person, as philosophical and everyday conceptions of the scientific picture of the world are freed from outdated Newtonian ideas and begin to take more note, both of the complexities of science as it really is and of the undeniable fact of our animal nature.[5]

1.0 NATURAL LIBERTARIANISM AND MINDED ANIMALISM

What is free will? What is practical agency? What is human personhood? And how are human free will, practical agency, and human personhood really possible in the natural world as it is correctly characterized by the modern natural sciences, especially physics, chemistry, biology, and cognitive neuroscience? Or more compactly put: given the truth of modern science, how is human free agency really possible? Let us call this *the freedom question*. In this book, I provide what I think is a rationally decisive and true answer to the freedom question.

But in order to formulate this answer clearly and distinctly, I will need some precisely-defined terminology. Correspondingly, I will say that *free will* is a conscious subject's power to choose or do what he or she wants to, or to refrain from so choosing or so doing, without preventative constraints and without internal or external compulsion, with at least causal responsibility; that *practical agency* is a conscious subject's power to choose or do things freely in the light of principles or reasons, including *moral* principles or reasons, on the basis of self-conscious processes of deliberation and decision; and that a *person* is an animal that is capable of free will, practical agency, and moral responsibility. So the freedom question is asking how human free will, human practical agency, and human personhood *really exist*, and how we *truly have them*, in the natural world as correctly described by modern science.

Or, on the contrary, given the truth of modern science, are we really nothing but "biochemical puppets"[6] or "moist robots,"[7] that is, nothing but natural automata, or natural machines, whose evolutionary and neurobiological mechanisms continually generate the cognitive illusion that we are free agents? If so, then we would be in an even worse cognitive place than Pinocchio, a wooden puppet who longed to be a real boy. We would be nothing but "meat puppets,"[8] dreaming that we are real human persons.

The issue being raised here, then, is how we should understand the implications of the modern natural sciences for our classical conception of ourselves as *rational and moral animals*, in the face of the possibility that we are under a serious cognitive illusion about this. Indeed, some contemporary philosophers even think that once we are liberated from this serious cognitive illusion, we will see finally clearly see that we are *nothing but* highly complex "biochemical puppets" *and* that "physics makes us free" in a deterministic, block universe.[9]

But quite frankly, in my opinion, any philosophical doctrine which holds

> (i) that we are nothing but biochemical puppets, no matter how highly complex and amazing these puppets are, and
> (ii) that "physics makes us free" in a deterministic, block universe,

is something straight out of George Orwell's 1984.[10] How politically expedient it would be for any 21st century equivalent of "Big Brother" to be able to convince us that our being nothing but highly complex decision-theoretic, deterministic automata and our being "free" are the same thing. On the contrary, then, it is a direct implication of my view that it is precisely those who believe and want to convince us that we are deterministic (or indeterministic) natural automata who are in the grip of a serious cognitive myth, not we who conceive of ourselves as purposive, living, essentially embodied, conscious, intentional, caring, really free rational and moral animals. Furthermore, critically resisting and truly liberating ourselves from this deeply-entrenched natural-mechanistic illusion about ourselves will have deep and radical ethical, sociocultural, and political implications. As D.R. Griffin and Helen Steward so rightly say in the fourth and sixth epigraphs for this chapter:

> There is an absolute contradiction between the freedom we all presuppose in practice and the implications of ideas that are widely accepted as established scientific fact. Philosophy has no higher calling than to try to resolve this contradiction at the heart of contemporary culture.

> The task [of understanding free will and agency] requires some reflection on the organizational principles of living creatures, for it is only through such reflection … that we can start to understand where the difference really lies between, on the one hand those things that are true agents, and, on the other, mere machines, entities that nothing will ever be up to, however impressive they may be…. I am exceedingly hopeful that the next few years will see the beginnings of a revolution in our conception of the human person, as philosophical and everyday conceptions of the scientific picture of the world are freed from outdated Newtonian ideas and begin to take more note, both of the complexities of science as it really is and of the undeniable fact of our animal nature.

Re-stated, then, my aim in this book is to provide a rationally decisive and true answer to the freedom question, with heavy-duty practical implications down the line,[11] by working out and defending what I call a *real metaphysics* of human free will, practical agency, and persons in a thoroughly nonideal natural and social world.

What is real metaphysics? As I pointed out in the Preface and General Introduction to THE RATIONAL HUMAN CONDITION (volume 1, part 1), there are two fundamental notions standing in the background here.

First, by *a veridical appearance* I mean anything *X* that appears as *F*, or appears *F*-ly, or appears to be *F*, to any or all rational human cognizers, *just insofar as, and precisely because, X is F*. For example, if I say "It appears that Sweetpea the cat is looking at me from her cat

cave," and what I say is indeed the case, as per this—

Figure 1. Sweetpea the cat is looking at me from her cat-cave.

or "It appears that 2 + 2 = 4," or "It appears that The Minimal Law of Non-Contradiction[12] applies universally," and again what I say is indeed the case, as per basic arithmetic and pure logic, then all the things I am talking about are veridical appearances.

Second, by the manifestly real world, I mean the world as it can veridically appear, or does veridically appear, to any or all rational human cognizers or agents.

Correspondingly, a statement (judgment, belief, proposition, meaningful sentence, etc.) is true if and only if what it states (means, says, etc.) is manifestly real. But the theory of truth is another story for another day.

In any case, granting those two bedrock notions, real metaphysics starts with the primitive, irreducible fact of purposive, living, essentially embodied, conscious, intentional, caring, rational and moral human experience in the manifestly real world, and then reverse-engineers its basic metaphysical theses and explanations in order to conform strictly to all and only what is phenomenologically self-evident in human experience.

By "phenomenologically self-evident" I mean this:

> A claim C is phenomenologically self-evident for a rational human subject S if and only if
>
> (i) S's belief in C relies on directly-given conscious or self-conscious manifest evidence about human experience, and

> (ii) C's denial is either logically or conceptually self-contradictory, really metaphysically impossible, or pragmatically self-stultifying for S.

This leads directly to what I call *the criterion of phenomenological adequacy for metaphysical theories*:

> A metaphysical theory is phenomenologically adequate if and only if that metaphysical theory is evidentially grounded on all and only phenomenologically self-evident theses.

Real metaphysics therefore rejects the idea of any theoretically fully meaningful, non-paradoxical ontic commitment or cognitive access to non-manifest, non-apparent, "really real" entities that are constituted by intrinsic non-relational properties—that is, "noumena" or "things-in-themselves."[13] Such entities are logically, conceptually, or "weakly metaphysically" possible, but strictly unknowable by minded animals like us, both as to their nature, and as to their actual existence or non-existence. In this sense, real metaphysics is *methodologically* eliminativist about noumena.

As such, real metaphysics sharply contrasts, for example, with *contemporary Analytic metaphysics*,[14] given the latter's commitments to noumenal realism, to Conceptualism about the nature of mental representation, to a heavy reliance on modal logic as providing direct insight into the ultimate structure of noumenal reality, usually combined with scientific essentialism, and to the unargued assumption that metaphysics is inherently non-normative and value-neutral. As an almost textbook-example of all this, in *Writing the Book of the World*, Ted Sider says:

> The central theme of this book is: realism about structure. The world has a distinguished structure, a privileged description. For a representation to be fully successful, truth is not enough; the representation must also use the right concepts, so that its conceptual structure matches reality's structure. There is an objectively correct way to "write the book of the world."… I connect structure to fundamentality. The joint-carving notions are the fundamental notions; a fact is fundamental when it is stated in joint-carving terms. A central task of metaphysics has always been to discern the ultimate or fundamental reality underlying the appearances. I think of this task as the investigation of reality's structure.[15]

Therefore, real metaphysics rejects all noumenal realist metaphysics, *especially including* contemporary Analytic metaphysics.

In the first half of the 20th century, the new and revolutionary anti-(neo)Kantian, anti-(neo)Hegelian philosophical programs were Gottlob Frege's and Bertrand Russell's *logicism*, G.E. Moore's *platonic atomism*, and the "linguistic turn" initiated by Wittgenstein's *Tractatus*, which yielded Russell's *logical atomism* and The Vienna Circle's *logical empiricism*, and finally its nemesis, W.V.O. Quine's critique of the analytic-synthetic distinction.[16] Logical empiricism also produced a domestic reaction, *ordinary language philosophy*. Powered by the work of H.P. Grice and Peter Strawson, ordinary language philosophy became *conceptual analysis*. In turn, Strawson created a new "connective," namely, holistic, version of conceptual analysis, that also constituted a "descriptive metaphysics."[17] Strawson's connective conceptual

analysis gradually fused with John Rawls's holistic method of "reflective equilibrium" and Noam Chomsky's psycholinguistic appeals to intuitions-as-evidence, and ultimately became the current *Standard Picture* of mainstream analytic philosophical methodology.[18]

Coexisting in mainstream contemporary philosophy, alongside the Standard Picture, is the classical Lockean idea that philosophy should be nothing but an "underlaborer" for the natural sciences, especially as this idea was developed in the second half of the 20th century by Quine and Wilfrid Sellars, as the reductive or eliminativist, physicalist, and scientistic doctrine of *scientific naturalism*, and again in the early 21st century in even more sophisticated versions, as "experimental philosophy" aka "X-Phi," and by Penelope Maddy, as the doctrine of *second philosophy*.[19]

From the standpoint of real metaphysics, what is wrong with scientific naturalism/X-Phi/second philosophy is its reduction or elimination of the primitive, irreducible fact of human experience. In turn, what is more subtle, but in some ways philosophically even more problematic, is when scientific naturalism/X-Phi/second philosophy takes on the seemingly philosophically more acceptable guise of *non-reductive physicalism*, and then quietly substitutes the fully ersatz causally inert, logically private or solipsistic, infallible, and ineffable epistemic and metaphysical pseudo-fact of *epiphenomenal qualia* for the primitive, irreducible, fully-causally-empowered fact of human experience.[20]

Again from the standpoint of real metaphysics, what is fundamentally wrong with the Standard Picture is its intellectualist, coherentist reliance on networks of potentially semantically empty, non-substantive *concepts*,[21] and above all, its avoidance of the sensible, essentially non-conceptual side of human experience and human cognition, which alone connects it directly to what is manifestly real.[22]

A further, and perhaps ultimately even more important problem with the Standard Picture, in view of the downstream deep and radical practical implications of real metaphysics, is the Standard Picture's striking lack of self-critical political awareness. For example, aside from Derek Parfit's *Reasons and Persons*, which I will critically address in chapter 7 below, certainly the other most brilliant and influential examples of philosophy done according to the Standard Picture are Rawls's *Theory of Justice* and Robert Nozick's *Anarchy, State, and Utopia*. But as many feminist and neo-Marxist critics have pointed out, Rawls's social-contractualist "veil of ignorance" methodology and Nozick's political libertarianism, alike, uncritically and questionably presuppose that late 20th century USA-style neo-Hobbesian, neo-Millian individualist political liberalism, or neoliberalism, are the *alpha* and *omega* of all serious political philosophy.

Real metaphysics is all about *the rational human condition*, and *not* about noumenal entities, fundamentally physical, essentially non-mental facts, or coherent networks of potentially semantically empty and/or "ideologically overdetermined" concepts.[23]

Re-re-stated now, another equivalent way of describing my aim in this book, is to say that it is an attempt to provide a rationally decisive and true answer to the freedom question, with heavy-duty practical implications down the line, by working out and defending a *contemporary Kantian real metaphysics* of human free will, practical agency, and persons in a thoroughly nonideal natural and social world. Like Kant, I hold that we directly experience practical agency, or what he calls "practical freedom," and also that practical agency requires metaphysically robust human free will or *deep freedom*—what he calls "transcendental freedom"—which really exists in the natural world. And I also fully agree with Kant that this kind of freedom "seems to be contrary to the law of nature, thus to all possible experience, and

so remains a problem." In Kantian terminology, I am trying to solve Kant's problem by developing a real metaphysics *according to which "transcendental freedom," and "practical freedom" or "autonomy," are themselves fully natural facts*. But in my own terminology, I am trying to solve Kant's problem by developing a real metaphysics according to which *deep freedom* and *practical agency* are themselves fully natural facts.

(From here on in, unless otherwise specified, and apart from a few places where I use the phrase "real metaphysics" for special emphasis, by using the term "metaphysics" I will always mean *real metaphysics* in the anti-noumenal-realist, anti-Analytic-metaphysics, anti-Standard Picture, anti-scientific naturalist/X-Phi/second philosophy, contemporary Kantian sense I have just spelled out.)

On the view I am proposing, our double capacity for free will and practical agency—which, for convenience, I will call *free agency*—is an irreducible fact. At the same time, our irreducible capacity for free agency does *not* exist over and above the rest of the physical world—it is categorically *not* a mysterious dualistic, extra-physical fact. On the contrary, it is a fully natural, biological and neurobiological fact—a natural fact of life. So the key to our free agency is not that we possess mysterious, non-natural, atemporal causal powers to choose or act in violation of the causal laws of nature. On the contrary, it is simply that, insofar as we are minded living organisms,

> (i) *all* of our intentional activities—by which I mean the things that we ourselves do, and do not merely happen to us—are inherently vital and non-mechanical,
> (ii) *some* of our intentional activities are naturally creative and authentically original, in just the way that a work created by a "human, all too human" artistic genius is authentically original, but not in any god-like or magical way, and
> (iii) *none* of the general causal laws of nature is ever violated by us.

In turn, these three natural facts, namely,

> (i*) vital, non-mechanical sourcehood
> (ii*) natural creativity, and
> (iii*) living in bounded natural open space (having non-equilibrium-thermodynamic "degrees of freedom"),

are all strongly supported by the self-evident, veridical phenomenology of rational human minded animal agency.[24]

More precisely, the realization of our capacity for free agency, exactly as we consciously experience it, is identical to *the form of our own lives* as rational human minded animals, that is, as life-forms of a certain kind. Indeed, our consciously experienced and thereby realized capacity for free agency exists fully *within* the causally efficacious biological and neurobiological facts that constitute the existence and specific character of our own organismic lives. And this is because our consciously experienced and thereby realized capacity for free agency is nothing more and nothing less than an *immanent structure* of those causally efficacious biological and neurobiological facts themselves. And these facts are, in turn, immanent structures of causally efficacious far-from-equilibrium, spatially orientable, temporally forward-directed, complex, self-organizing thermodynamic processes. According to this non-reductive, immanent structuralist, dynamicist metaphysical conception, then, free

agency is an immanent structure of conscious, intentional mind; conscious, intentional mind is an immanent structure of organismic life; and organismic life is an immanent structure of spatiotemporally asymmetric matter and/or energy flows. Each more complex structure is metaphysically continuous with, and embeds, all of the less complex structures. Here is a simple diagram of the basic metaphysical continuities and structural embeddings:

> free agency→conscious, intentional mind→organismic life→asymmetric matter/energy flows

In view of this metaphysical conception, to borrow an apt phrase from the later Wittgenstein, my own free agency is just my own "form of life," and free agency, as such, grows naturally in certain minded animal species or life-forms. Correspondingly, freedom grows naturally and evolves in certain species of minded animals, including the human species, precisely because minds like ours grow naturally and evolve in certain species of animals, including the human species.[25] This thesis, which I will call *The Freedom-in-Life Thesis*, is a central part of the doctrine I call *Natural Libertarianism*. Freedom is dynamically inherent in and dynamically emerges from mind, and mind is dynamically inherent in and dynamically emerges from life; hence freedom is dynamically inherent in and dynamically emerges from life. Life is dynamically inherent in and dynamically emerges from asymmetric matter and/or energy flows. So freedom, mind, and life are all dynamically inherent in and dynamically emerge from asymmetric matter and/or energy flows.[26]

But the "inherent in" here is *not* a reductive "inherent in," and the "emerges from" is *not* a supervenient or dualistic "emerges from." Freedom, mind, and life are all "dynamically inherent in" and "dynamically emerge from" asymmetric matter and/or energy flows *only* in the same basic metaphysical sense that transfinite numbers, complex numbers, and real numbers are all non-dynamically "inherent in" and non-dynamically "emerge from" the self-same dense mathematical structural continuum that also fully embeds the progressively simpler sub-structures consisting of the rational numbers or fractions, the integers, and the natural numbers. Inside this structural continuum, the transfinite, complex, and real numbers are *not* Turing-computable functions of the natural numbers, integers, or rational numbers. And yet, given the existence of the natural numbers, integers, and rational numbers, necessarily the real, complex, and transfinite numbers are all potentially *there*, most of them existing *between* the rationals, integers, and naturals, holding them all *together* within progressively more complex forms of unity, waiting to emerge by mathematical discovery.

In short, the dynamic inherence of freedom, mind, and life is nothing more and nothing less than causally efficacious *immanent structural inherence*, and the dynamic emergence of freedom, mind, and life is nothing more and nothing less than causally efficacious *immanent structural emergence*. So just as it would be mathematically absurd to try to reduce transfinite, complex, or real numbers to recursive functions of the rationals, integers, or naturals, so too it would be metaphysically absurd to try to reduce freedom, mind, and life to Conservation-Laws-governed causal functions of inert, mechanical matter or material processes. Yet that is precisely what defenders of *reductive physicalism* try to do. And just as it would be mathematically absurd to think of the transfinite, complex, or real numbers as existing "over and above" the rationals, integers, or naturals—on the contrary, the number designated by '2', for example, is a distinct position or role *inside* the system of natural numbers, *inside* the system of positive integers, *inside* the system of rational numbers, *inside* the system of real numbers,

inside the system of complex numbers, and *inside* the system of of transfinite cardinals—so too it would be metaphysically absurd to think of freedom, mind, and life as existing "over and above" asymmetric matter/energy flows. Yet this is precisely what *non-reductive physicalists* and *ontological dualists* about freedom, mind, and life try to do. I will have much more to say about these fundamental metaphysical points later.

If Natural Libertarianism is correct, then it also directly entails biologically-oriented theories of what I call *real* personhood and *real* personal identity, which I call, collectively, *Minded Animalism*. According to Minded Animalism, you and I and the folks living next door are all real persons who are each literally identical—by which I mean numerically identical, or token-identical, and also personally identical—one-to-one, with the complete, finite, and unique *life* of some individual minded animal in the class of all minded animals. Then real personal identity is not an identity relation between a mind and a mind, or between a body and a body, or even between an animal and an animal. On the contrary, real personal identity is an identity relation between an animal's *life-process* and an animal's *life-process*.[27]

What kinds of animal lives am I talking about? Some animals are minded, like us, and some animals are not minded, for example, human babies born with anencephaly, human beings in persistent vegetative states, and perhaps also insects and reptiles. In any case, by a *minded animal*, I mean any living organism with inherent capacities or powers for:

> (i) *consciousness*, that is, a capacity for embodied subjective experience,[28]
> (ii) *intentionality*, that is, a capacity for conscious mental representation and mental directedness to objects, events, processes, facts, acts, other animals, or the subject herself (so in general, a capacity for mental directedness to *intentional targets*), and also for
> (iii) *caring*, a capacity for conscious affect, desiring, and emotion, whether directed to objects, events, processes, facts, acts, other animals, or the subject herself.

Over and above consciousness, intentionality, and caring, in *some but not all* minded animals, there is also a further inherent capacity for

> (iv) *rationality*, that is, a capacity for self-conscious thinking according to principles, responsiveness to reasons, and reasons-seeking, hence poised for justification, whether logical thinking (including inference and theory-construction) or practical thinking (including deliberation and decision-making).

In chapters 6 and 7, I will argue that it is the fourth capacity (rationality), naturally built on top of the other three capacities (consciousness, intentionality, and caring), that sufficiently constitute the real personhood of a minded animal. Now some minded animals are human, like us, and some minded animals are non-human, and also, arguably, are *rational* minded animals and real persons—for example, Great apes (by which I mean non-human members of the biological family Hominidae, including bonobos, chimpanzees, gorillas, and orangutans) and dolphins. Furthermore, some real persons, like us, who are capable of reading and understanding these words, are fully self-conscious and reflective, fully capable of doing creative art, science, and philosophy, fully capable of practical agency, and fully capable being morally responsible for our freely-willed choices and acts, for better or worse. Perhaps some

of these "higher-level" or "Kantian" real persons are also non-human! But at the very least, we do know some human ones intimately: ourselves. All human beings are animals, but some human animals are not minded animals—for example, again, anencephalic babies, normal human fetuses prior to 25 weeks after conception, and humans in persistent vegetative states. Moreover, all rational animals, including of course rational human animals, like us, are minded animals, but some human animals are not rational, including of course all the non-minded human animals, but also the incurably insane. And all rational animals, including of course rational human animals, are real persons, but not all minded animals, including some human animals, are real persons, including normal cats, dogs, horses, and mice, human beings in the final stages of Alzheimer's disease, and again the incurably insane.

So I am saying that free agency is an inherently "onboard" biological and neurobiological guidance system in any rational minded animal, or real person, including of course any rational human minded animal, or real human person, like us. Even more precisely stated, I am saying that free agency is an intra-organismic, desire-based and emotive, more or less rationally principled, more or less wholehearted, active guidance system that can effectively initiate, coherently maintain, and intimately control the movements of our own living animal bodies under favorable local environmental conditions. But I am *also* saying that some non-rational minded animals, including some human animals and some non-human animals, are possessed of the power of what I call *free volition*,[29] by virtue of their *also* possessing an intra-organismic biological and neurobiological active guidance system that can effectively initiate, coherently maintain, and intimately control the movements of their own living organismic bodies under favorable local environmental conditions.

Correspondingly, free agency and free volition alike are primitively guided by what contemporary philosophers of mind and knowledge call "non-conceptual mental content," and more specifically, it is guided by what I call *essentially non-conceptual content*.[30] So the basic cognitive capacity for knowing our own free choices and basic acts, as trying-initiated and actively-guided intentional body-movements, is the capacity for *essentially non-conceptual cognition*. That is the key link between volume 5 of THE RATIONAL HUMAN CONDITION, *Cognition, Content, and the A Priori*, and the present book, volume 2. Freedom and essentially non-conceptual cognition are vital powers that extend significantly more widely across the domain of minded animals than do specifically *rational human* free agency and specifically *human* real personhood.

Nevertheless, specifically rational human free agency and specifically human real personhood are my primary philosophical targets in this book. This is not only because I want to answer the freedom question, but also because I want to use them to provide the basic metaphysical framework for the new version of contemporary Kantian moral theory that I develop in volume 3 of THE RATIONAL HUMAN CONDITION, *Kantian Ethics and Human Existence*,[31] and, in turn, for the new version of contemporary Kantian philosophy of religion/philosophical theology and political philosophy that I develop in volume 4 of THE RATIONAL HUMAN, *Kant, Agnosticism, and Anarchism*.[32] At the same time, rational human free agency and human real personhood alike, as *epistemic* facts, are grounded on what, in *Cognition, Content, and the A Priori*, I call *non-conceptual knowledge*; and, as *ontological* facts, they are also both metaphysically grounded on what I have called *essentially embodied agency*, or *minded animal agency*.[33]

Natural Libertarianism therefore builds directly on the unified account of consciousness, the mind-body relation, mental causation, and intentional action that Michelle Maiese and I

worked out in *Embodied Minds in Action*—The Essential Embodiment Theory. Here are the six central theses of that theory:

> *(1) The Essential Embodiment Thesis*: Creatures with conscious, intentional minds are necessarily and completely neurobiologically embodied.
> *(2) The Essentially Embodied Agency Thesis*: Basic acts (for example, raising one's arm) are intentional body movements caused by an essentially embodied mind's synchronous trying to make those very movements and its active guidance of them.
> *(3) The Emotive Causation Thesis*: Trying and its active guidance, as the cause of basic intentional actions, is primarily a pre-reflective, desire-based emotive mental activity and only derivatively a self-conscious or self-reflective, deliberative intellectual mental activity.
> *(4) The Mind-Body Animalism Thesis*: The fundamental mental properties of conscious, intentional minds are (a) non-logically or strongly metaphysically (namely, synthetically) a priori necessarily reciprocally intrinsically connected to corresponding fundamental physical properties in a living animal's body (mental-physical property fusion), and (b) irreducible truly global or inherently dominating intrinsic structures of motile, suitably neurobiologically complex, egocentrically-centered and spatially-oriented, thermodynamically irreversible living organisms (neo-Aristotelian hylomorphism).
> *(5) The Dynamic Emergence Thesis*: The natural world itself is neither fundamentally physical nor fundamentally mental, but is instead essentially a causal-dynamic totality of forces, processes, and patterned movements and changes in real space and real time, all of which exemplify fundamental physical properties (for example, molecular, atomic, and quantum properties). Some but not all of those physical events also exemplify irreducible biological properties (for example, being a living organism), and some but not all of those biological events also exemplify irreducible fundamental mental properties (for example, consciousness or intentionality). And both biological properties and fundamental mental properties are dynamically emergent properties of those events.
> *(6) The Intentional Causation Thesis*: A mental cause is an event or process involving both consciousness and intentionality, such that it is a necessary proper part of a nomologically jointly sufficient essentially mental-and-physical cause of intentional body movements. In so doing, it is a dynamically emergent structuring cause of those movements. Then, under the appropriate endogenous and exogenous conditions, by virtue of synchronous trying and its active guidance, conscious, intentional essentially embodied minds are mental causes of basic acts from their inception in neurobiological processes to their completion in overt intentional body movements.

In this way, The Essential Embodiment Theory says that our dynamically emergent, irreducible, sentient and sapient minds are also necessarily interdependent with our own living organismic animal bodies and not essentially distinct from them; that we are far-from-equilibrium, asymmetric, complex, self-organizing thermodynamic systems; that we act by intentionally moving our bodies by means of our desire-based emotions and trying; and that our conscious, intentional, caring, and rational necessarily and completely neurobiologically

embodied minds are basically causally efficacious precisely because they are metaphysically continuous with our biological lives, and life is basically causally efficacious in physical nature. The simple upshots of The Essential Embodiment Theory, then, are:

> (i) in thinking about the mind-body problem we should decisively replace the early modern Cartesian and Newtonian *ghost-in-the-machine* metaphysics with a post-Cartesian and post-Newtonian but also at the same time neo-Aristotelian *immanent-structure-in-the-non-equilibrium-thermodynamics* metaphysics, and
> (ii) the irreducible conscious, intentional, caring minds of cognizers and agents grow naturally in suitably complex living organisms, as irreducible, non-dualistic, non-supervenient, asymmetric thermodynamic structures *of* those organisms.

For obvious reasons of space-economy, and also simply in order to avoid redundancy across books, I will not undertake to re-present or re-defend The Essential Embodiment Theory here. At the same time, I will presuppose the truth of its six central theses, as having already been sufficiently elaborated and justified in *Embodied Minds in Action*, in order to use them for my present philosophical purposes. To orient critical readers looking specifically for that elaboration and justification, however, I will also indicate the relevant corresponding chapters or sections of *Embodied Minds in Action* whenever I am drawing or relying on that material.

Most simply described, Natural Libertarianism is The Essential Embodiment Theory insofar as it has been explicitly extended to free volition and minded animal agency in general, and also to free agency, namely, to free will and practical agency, namely, to rational human minded animal agency, in particular. Other things being equal, and within certain natural limits, as rational human animals, we can move our own bodies just *because* we want to, and just *as* we want to, more or less from rational principles, and more or less wholeheartedly. We are then free agents by virtue of the fully natural fact that each rational human minded animal like us is a unique life-form that self-determines her own unique form of life, in a more or less principled and wholehearted way, under or within various larger-scale and smaller-scale pre-existing physical, chemical, biological, and neurobiological causal-law-determined constraints or parameters, which necessarily in turn are also natural dynamic "enablers" for making our choices and actions causally possible. So our *agential* or *agentive* autonomy[34] is nothing more and nothing less than a suitably complex natural development and outgrowth of our causal-nomological autonomy and our biological autonomy.

1.1 INCOMPATIBILISTIC COMPATIBILISM

Here, again, is the same basic idea behind Natural Libertarianism, but this time formulated in the terminology of the contemporary philosophical debate about free will.[35] The classical thesis of *Universal Natural Determinism* says that all present and future spacetime events, including all the things we specifically choose and do, are necessitated by all the settled facts about the past, together with all the general causal laws of nature. So if Universal Natural Determinism is true, then either whatever I am choosing or doing now is directly necessitated by The Big Bang (I will call this *distal* determination), or else at the very least there is a more spatiotemporally local deterministic process or state of the physical world, which, together with the settled facts about the past and the laws of nature, necessitates what I choose or do now (I

will call this *proximal* determination). Otherwise put, if Universal Natural Determinism is true, then either The Big Bang causally flows right through me (= distal determination), or else the more local physical environment causally flows right through me (= proximal determination). In either case, if Universal Natural Determinism is true, then all the events in my life, now and henceforth until I die, are lawfully causally necessitated by the settled facts about the past—really, I am a deterministic natural automaton.

Now the classical thesis of *Compatibilism*[36] says that free will and Universal Natural Determinism are mutually consistent: that is, it is logically or metaphysically possible that at least *some* spacetime events are both free and also determined in a world in which all spacetime events are determined, and also logically or metaphysically possible that *all* free spacetime events are also determined in a world in which all spacetime events are determined. Correspondingly, the classical thesis of *Incompatibilism*[37] says that free will and Universal Natural Determinism are mutually inconsistent: that is, it is logically or metaphysically impossible that *any* spacetime events are both free and determined.

Many contemporary philosophers are Compatibilists, and some contemporary philosophers are Incompatibilists. But Natural Libertarianism is neither strictly Compatibilist nor strictly Incompatibilist, and it thereby constitutes what David Hodgson aptly calls a "third alternative" to this all-too-familiar and seemingly exhaustive dichotomy, which I dub *Incompatibilistic Compatibilism*. In a nutshell, according to the Incompatibilistic Compatibilism of Natural Libertarianism, free will and Natural Determinism are *locally incompatible* but also *non-locally compatible*.

To classical Incompatibilists and Compatibilists alike, this may well seem absurd. How can Natural Libertarianism consistently propose the theses of both *Local Incompatibilism* and also *Non-Local Compatibilism*? Or otherwise put, how can one consistently be an incompatibilistic compatibilist? That, surely, is like wooden iron, or round squareness. But in fact there is no incoherence whatsoever between Local Incompatibilism and Non-Local Compatibilism. More precisely, then, here is how one can consistently be that strange philosophical beast, an "incompatibilistic compatibilist."

On the one hand, Local Incompatibilism says that neither the existence nor the specific character of the particular actual conscious and intentional choices or acts of rational or non-rational human or non-human minded animals, as living organisms having a certain kind of non-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic structure inherent in their individual biological and neurobiological lives, now or in the future, is itself necessarily or constitutively determined by all the general laws of nature together with all the settled facts about the past. For if this necessary or constitutive determination had been true, then there would have been be an inherently *non-agential* production of the minded animal's choices or acts either by means of distal determination (that is, The Big Bang would actually be causally responsible) or else by means of proximal determination (that is, the more local physical environment would actually be causally responsible). But whether she were distally or proximally determined, then the rational or non-rational human or non-human minded animal would not be a free agent with respect to those choices or acts, precisely because she would not even be an *agent*. For if either The Big Bang or the more local physical environment were to be actually causally responsible for what she chooses and does, then she would not be causally responsible for what she chooses and does, they would not be "up to her," she would not be an "ultimate source" of those choices or acts, and as a consequence, in the special case of rational agents, she would not be morally responsible either. For she would

be a deterministic natural automaton, or natural machine, with respect to those choices or acts. Indeed, it would just as if some other very powerful agency or person—for example, an evil super-scientist like the weirdly jolly Communist behavioral psychologist in John Frankenheimer's brilliant 1962 paranoid thriller, *The Manchurian Candidate*—had overwhelmingly compelled, forced, or manipulated her into making those choices or carrying out those acts.[38]

On the other hand, however, Non-Local Compatibilism says that the actual existence of free volition or free will in minded animals, especially including rational animals, including of course all rational human animals, is not only logically and metaphysically consistent with but also metaphysically requires the existence of a complete set of general causal laws obtaining in the rest of the physical world. This complete set of general causal laws, in turn, together with all the settled facts about the past, causally necessitate the existence and specific character of a great many physical processes in the present and the future. But this causal necessitation operates independently of *all those thermodynamic processes that constitute any rational or non-rational human or non-human minded animal's own biological and neurobiological life*. At the same time, however, this causal necessitation remains a necessary background condition of those very same causal-necessitation-independent processes that constitute any rational or non-rational human or non-human minded animal's own biological and neurobiological life. For example, in order for me to hop up and down a little as a basic intentional act, the law of the acceleration of falling bodies due to gravity must remain fixed so that I can use and exploit it for my own purposes. But this gravitational law, together with all the settled facts about the past, does not causally necessitate my act of hopping. There is still some non-deterministic "natural open space" left over, within which I can purposively operate. And this generalizes across all deterministic general causal laws, and all the settled facts about the past. In other words, the causal necessitation is necessary but *not* sufficient for the existence and specific character of the life-processes of those minded animals.

Obviously Local Incompatibilism is not precisely the same thesis as classical or strict Incompatibilism, just as Non-Local Compatibilism is not precisely the same thesis as classical or strict Compatibilism. Classical or strict Incompatibilism says that free will and Natural Determinism are logically or metaphysically inconsistent always and everywhere, and in every logically or metaphysically possible world. But Local Incompatibilism says, by contrast, that in this actual world, free agency in rational or non-rational human or non-human minded animals is metaphysically inconsistent with deterministic physical processes *only insofar as* the existence and specific character of the self-organizing thermodynamic biological and neurobiological processes that are themselves identical with the free agency of rational or non-rational human or non-human minded animals *are actually concerned*. So Local Incompatibilism is a specially restricted version of Incompatibilism. And because it is specially restricted in this way, Local Incompatibilism is also perfectly logically and metaphysically consistent with a correspondingly specially restricted version of Compatibilism, namely, Non-Local Compatibilism.

Correspondingly, then, unlike classical or strict Compatibilism, which says that free will and Natural Determinism are logically or metaphysically consistent always and everywhere, and in every logically and metaphysically possible world, by contrast Non-Local Compatibilism says that in this actual world there is a complete set of general causal laws, which, together with all the settled facts about the past, causally necessitate the existence and specific character of a great many physical processes in the present and the future, precisely

insofar as they are *not* self-organizing thermodynamic systems, including the lives of minded animal free agents. And it also says that this complete set of general causal laws and its corresponding set of deterministic processes are actually presupposed by the existence and specific character of the self-organizing thermodynamic biological and neurobiological processes, including those that are themselves identical with the free agency of minded animals. "Non-Local" thus means "across the actual world, other than those parts of the world that are self-organizing thermodynamic systems, including the lives of minded animal free agents" and not "across every logically and metaphysically possible world." So Non-Local Compatibilism is a specially restricted version of Compatibilism. And because it is specially restricted in this way, Non-Local Compatibilism is also perfectly logically and metaphysically consistent with a correspondingly specially restricted version of Incompatibilism, namely, Local Incompatibilism.

So I am saying that whereas an *un*restricted Incompatibilism and an *un*restricted Compatibilism are logical contradictories, or at the very least logical contraries, nevertheless a specially restricted Incompatibilism is perfectly consistent with a specially restricted Compatibilism. Of course you may still be wondering why I am, seemingly, splitting logical and metaphysical hairs! But, at least by my intention, this is not philosophical hair-splitting. I am, in effect and in truth, trying to explain how what Wilfrid Sellars so aptly called *The Scientific Image* of rational human minded animals in-the-world is capable of logically and metaphysically cohabiting with what Sellars equally aptly called *The Manifest Image* of rational human minded animals in-the-world.[39] That is, I am trying to explain how our and other minded animals' free agency is both locally inconsistent with Natural Determinism and also non-locally consistent with Natural Determinism. Or in still other words, I am trying to explain how nature's worldwide causal-nomological mechanism occurring *beyond* and *outside* self-organizing thermodynamic systems like us can also coherently co-exist in the selfsame actual world as nature's localized rational or non-rational human or non-human minded animal free agency occuring *within* and *inside* self-organizing thermodynamic systems like us.

The fundamental point is that the right conception of the physical natural world metaphysically folds irreducible phenomenological-structural and teleological-structural facts, including life, consciousness, intentionality, caring, and free agency *directly into nature*, and thereby necessarily includes a representation of the inherent limits of a universal causal-mechanistic conception of physical nature. Again, it would be as if one were to try to explain transfinite, complex, and real numbers solely in terms of rationals, integers, and natural numbers, together with all the computable or recursive functions over them. Universal *formal* mechanism—the putative reduction of all of mathematics and logic to computability theory—manifestly does not work: so, since human mathematical knowing and logical knowing would have to be among the processes that are nothing but causal-mechanical processes, then why should we think that universal *causal-nomological* mechanism would work in metaphysics?[40] The "right conception of the physical natural world" that I am sketching here, and directly opposing to any universal causal-mechanistic view of physical nature, is what I call *liberal naturalism*.

In order for what I have just written to make fully clear and distinct sense to you, I also need to provide working definitions and brief explications of the basic notions I have already been deploying.[41] *Dynamic systems* are unified collections of material elements in rule-governed or patterned motion. In connection with dynamic systems, *complexity* is the fact that the causally efficacious exchange of matter and/or energy between a dynamic system and its

local natural environment does not remain constant, or fluctuates. Given complexity, then dynamic systems with identical, or virtually identical, initial conditions, may diverge radically over time. *Thermodynamic systems* necessarily involve energy (and degrees of energetic activity, heat), along with matter. Thermodynamic systems for which the formal structures of matter and/or energy remain the same, or essentially the same, over time, are *equilibrium, or near-equilibrium, time-reversible* systems. *Entropy* is a function of the state of a thermodynamic system that expresses the increasing amount of so-called "disorder" or "heat death" in a system, but less conceptually misleadingly, can be thought of as the increasing amount of *structural simplicity or unformity* in a system, which rises monotonically to a maximum at equilibrium conditions. Here energy is entirely potential, without actualization or entropic motion. So equilibrium or near-equilibrium, time-reversible thermodynamic systems do not (significantly) increase entropy. By contrast, thermodynamic systems for which the structures of matter and/or energy change over time, and are temporally irreversible in that they (significantly) increase entropy but do not reach a maximum of entropy, are *far-from-equilibrium* systems. *Self-organizing* complex thermodynamic systems, in turn, are far-from-equilibrium, temporally irreversible thermodynamic systems that also have *dissipative structure* and *natural purposiveness or natural teleology*. A dissipative structure is how the increasing amount of entropy in a complex thermodynamic system is absorbed and dispersed (hence "dissipated") by the systematic re-introduction of matter and/or energy into the system, via a non-static causal balance between the inner states of the system and its surrounding natural environment. And natural purposiveness or natural teleology is how a far-from-equilibrium, temporally irreversible, complex thermodynamic system with dissipative structure self-generates forms or patterns of order that determine its own causal powers, and in turn place constraints on the later collective behaviors, effects, and outputs of the whole system, in order to maintain itself. The paradigmatic or prime example of a self-organizing complex thermodynamic system is a living organism—although not every self-organizing system is itself an organism.

It is important to note here that non-equilibrium thermodynamics, aka "complex systems dynamics," as such, is nothing more and nothing less than an extremely interesting theory in mathematical natural science—indeed, a specifically post-Newtonian, non-mechanistic theory in mathematical natural science—and not a metaphysical theory. As such, non-equilibrium thermodynamics does not, in and of itself, entail the denial of Universal Natural Determinism. It would *also* be incorrect to say, however, that non-equilibrium thermodynamics is even metaphysically consistent with Universal Natural Determinism, although this is often assumed. That consistency would need to be shown by independent metaphysical arguments. My point is just that non-equilibrium thermodynamics, in and of itself, is metaphysically neutral, and significantly open to different metaphysical interpretations. And I am, indeed, offering a specific metaphysical interpretation of it, in terms of Natural Libertarianism. So non-equilibrium thermodynamics on its own does not entail Natural Libertarianism, although it remains perfectly consistent with it.

In any case, now granting me for the purposes of argument, the mathematical natural scientific idea of a self-organizing thermodynamic system as a working hypothesis under a specific metaphysical interpretation of it (that is, under Natural Libertarianism), then how can Local Incompatibilism and Non-Local Compatibilism both be true in the light of this working hypothesis, and under that specific metaphysical interpretation of it?

To begin with, in my opinion, we need to recognize that there is a fundamental modal distinction between

> (i) mere *consistency with* natural laws, and
> (ii) strict entailment or other necessitation by natural laws,

or more precisely, that there is a fundamental modal distinction between

> (i) a natural event's being *merely in conformity with*, in the sense of merely being the case along with and not in any violation of, all the general causal laws of nature together with all the settled facts about the past, and
> (ii) a natural event's being *strictly entailed or otherwise necessitated by*, in the sense of being sufficiently yielded by, all the general causal laws of nature together with all the settled facts about the past.

This crucial contrast, in turn, is a generalization of Kant's well-known modal distinction between

> (i) acting merely *according to* a moral principle or rule, and
> (ii) acting strictly *from* or *for the sake of* a moral principle or rule (*GMM* 4: 397-398).[42]

When this modal distinction is fully generalized beyond intentional action and deontological (that is, duty-sensitive, choice-involving) contexts, however, to contexts involving physical behaviors, functions, and operations of any kind, and indeed to contexts involving necessitation and rule-following of any kind—whether deontological, causal, mathematical, or logical—then this is the same as the fully comprehensive conceptual and metaphysical modal distinction between:

> (i) an activity's mere *conformity to a law* (or rule), and
> (ii) an activity's strict *governance by a law* (or rule).

Now an activity's mere conformity to a law or rule places a constraint or parameter on all that activity's behaviors, functions, and operations, thus *enabling* that activity's behaviors, functions, and operations, but without also thereby strictly *entailing* or otherwise necessitating either the existence or specific character of that activity's behaviors, functions, and operations. Let me now apply this thought specifically to the metaphysics of physical nature.

According to thethought I just describeded, all the general causal laws of nature together with all the settled facts about the past can, in at least some cases, be merely *constrainers* or *parameters* of physical activity in the present and the future, and indeed also be *causal enablers* of physical activity in the present and the future, but without also thereby being *strict entailers or necessitators* of that present or future physical activity. In chapter 2, I will propose that the correct physical interpretation of this condition of being merely consistent with all the general causal laws of nature is *conserved total quantity of matter and/or energy from natural cause to natural effect*. The conservation of total quantity of matter and/or energy in causal interactions, in turn, while it guarantees causal efficacy, causal closure, and physicality in a thermodynamic

system, and while its specific quantitative character is computable on a universal Turing-machine from all the settled energy facts about the past together with the general causal laws of nature, does *not* itself necessarily determine certain essential properties of thermodynamic systems. For example, it does *not* necessarily determine spatial intrinsic directionality or orientability with egocentric centering (indexicality) that makes topological asymmetries like enantiomorphy (aka "incongruent counterparts") really possible; it does *not* necessarily determine the direction of time and other temporal asymmetries; it does *not* necessarily determine facts about quantum complementarity or entanglement, quantum singularities like black holes, or the quantum collapse of the wave function; it does *not* necessarily determine chaotic behavior in far-from-equilibrium thermodynamic systems, or complex dynamic self-organization, including organismic life; and above all, it does *not* necessarily determine consciousness and intentionality, caring, free agency, or the unfolding of persons' lives. According to the metaphysical conception I am proposing, the conserved total quantity of matter and/or energy in causal interactions is the *cosmic minimal skeleton* of nature—in the same sense in which the natural numbers are the *mathematical minimal skeleton* of number theory as a designated or "identified" sub-structure within all number theories. But all the really interesting natural activity occurs in the non-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic *fleshing out* of that cosmic minimal skeleton.

If this metaphysical thought istrue, then to that extent physical nature will necessarily have a certain canalized or channelled amount of *non*-deterministic open texture, namely, a certain amount of "natural open space," which is also *not* itself merely *in*deterministic open texture, namely, natural chance or natural randomness. Again according to this metaphysical thought, all those thermodynamic systems whose causal characteristics and powers really *are* strictly entailed or otherwise necessitated by all the general causal laws together with all the settled facts about the past, are the same as all those natural systems for which conserved total quantity of matter and/or energy itself suffices for the necessary determination of all that system's basic natural properties, all of which, in turn, are Turing-computable from all the settled matter-and/or-energy-quantity-facts about the past together with the general laws of nature. Hence these systems are all nothing but *natural automata* or *natural machines*. But by sharp contrast, all those thermodynamic systems whose causal characteristics and powers are merely constrained or parametrized by all the general causal laws together with all the settled facts about the past, and are also causally enabled by those laws and facts, and therefore for which conserved total quantity of matter and/or energy, together with all the settled quantity-of-matter-and/or-energy facts about the past, together with the general causal laws of nature, does *not* suffice for the necessary determination of all their basic natural properties, are in fact far-from-equilibrium, spatially orientable, temporally irreversible, complex *self-organizing* thermodynamic systems. This is precisely because such systems are capable of exploiting the channelled non-deterministic, open-textured, open-spaced, and causally enabling, yet still non-indeterministic, non-chancy, non-random aspects of physical nature, *for their own self-produced (aka "autopoietic"*[43]*) ends*. In short, such systems are all naturally *purposive* or naturally *teleological* systems.

So, to summarize, here is the metaphysical picture I am proposing, according to which nature as a whole is categorically divided into three essentially different kinds of thermodynamic systems:

> *(1) Deterministic systems:* Thermodynamic systems whose behaviors, functions, and operations are nomologically necessitated in the present and the future by all settled facts about the past together with all general causal laws of nature, especially including the Conservation Laws, that operate under equilibrium, or near-equilibrium conditions, and are temporally reversible, are *deterministic natural automata*. Such systems entail the existence of closed futures, together with natural mechanism.
>
> *(2) Indeterministic systems:* Thermodynamic systems whose behaviors, functions, and operations are *not* nomologically necessitated in the present and the future by all settled facts about the past together with all general causal laws of nature, especially including the Conservation Laws, that may operate under far-from-equilibrium conditions and be temporally irreversible, *yet still occur now and in the future within a fixed probability space that is nomologically necessitated by all settled facts about the past together with all the statistical/stochastic general causal laws of nature*, are *indeterministic natural automata*. Such systems entail the existence of open futures, together with natural mechanism.
>
> *(3) Non-deterministic, non-indeterministic (namely, naturally purposive or naturally teleological) systems:* Thermodynamic systems whose behaviors, functions, and operations are constrained by and consistent with, but whose causal characteristics and powers are not nomologically necessitated either in the present or the future by, all settled facts about the past together with all general causal laws of nature, especially including the Conservation Laws and all the statistical/stochastic laws, that always operate under far-from-equilibrium conditions and are spatially orientable and temporally irreversible, are *self-organizing systems in natural open space*. Such systems entail the existence of open futures, without natural mechanism.

In other words, naturally mechanistic systems can be *either* deterministic *or* indeterministic, and more generally are all and only those thermodynamic systems for which all their basic natural properties, their causal characteristics and powers, are nothing but conservation-of-quantity-of-matter-and/or-energy facts, that are necessarily determined by and Turing-computable from all the settled energy-quantity facts about the past together with the general causal laws of nature. Some of these naturally mechanistic thermodynamic systems, and in particular some of the indeterministic ones, may *also* be far-from-equilibrium, temporally irreversible, and complex. But far-from-equilibrium, temporally irreversible, complex thermodynamic systems that are non-deterministic yet also *non-indeterministic*—namely, self-organizing, naturally purposive, naturally teleological systems—are therefore *not* natural mechanisms.

I especially emphasize this point, because even amongst the foremost theorists of non-equilibrium, asymmetric, complex systems thermodynamics, there has been a certain tendency to confuse non-deterministic, non-indeterministic, self-organizing, naturally purposive, naturally teleological (hence, *non-mechanistic*) thermodynamic systems with *naturally mechanized indeterministic thermodynamic systems*.[44] For example, Ilya Prigogine entitled his brilliant, breakthrough book in this area, *The End of Certainty*, thereby unfortunately running together the importantly distinct concepts of:

> (i) non-mechanistic non-determinism, and

> (ii) mechanistic indeterminism.

But as Prigogine himself recognized, organismic life is *not* just a highly involved kind of natural dice-playing. So really, and far less misleadingly, his breakthrough book should have been called *The End of Mechanism*.

I should also mention here explicitly, flagging this issue for further discussion in chapters 2 and 3 below in the context of the philosophy of biology, that I am thinking of causal laws within a neo-Aristotelian and contemporary Kantian metaphysical structuralist framework[45] as large-scale, multiply embedded, sets of immanent structures in actual manifest nature, and *not* in one of the more standard ways, as

> either (i) mere logical or conceptual entailment-relations between event-instances of natural universals (aka "The Armstrong-Tooley-Dretske conception" of causal laws),
> or (ii) mere natural regularities between events, involving counterfactual influence or probabilistic determination (aka "The Humean conception" of causal laws).[46]

In this metaphysical structuralist framework, the system of general deterministic or indeterministic causal laws, and especially the Conservation Laws, constitutes an intrinsic structural minimal matter and/or energy-grid, a designated or "identified" sub-structure, whose immanence in total physical nature yields the cosmic minimal skeleton of total-quantity-of-matter-and/or-energy conservation in all causally efficacious natural activity.

What is most important in the present context, however, is just my neo-Aristotelian and contemporary Kantian metaphysical thought that causal laws are *hylomorphically* related to natural processes as intrinsic relational form is to formed matter, thereby imposing various modal conditions on the spatiotemporal relationships between natural singular events; and that therefore causal laws are not in any way *extrinsic* to those natural singular events and the processes constituted by their relationships. If this metaphysical thought is correct, then causal laws are nothing more and nothing less than multiply-embedded, immanent structural properties of *all* thermodynamic systems, but especially including the far-from-equilibrium, spatially orientable, temporally irreversible, complex thermodynamics of self-organizing natural systems.

Moreover, unfolding my line of metaphysical thinking a bit more, not every far-from-equilibrium, asymmetric, complex, self-organizing thermodynamic system—for example, the roiling movements of boiling water, or a weather system—is capable of *successfully* or *unsuccessfully* exploiting the open-textured, open-spaced, enabling, law-governed internal relational aspects of the material world, in the sense of being able to sustain itself, and re-energize itself, in an appropriate interactive causal balance with its local environment, over a significant amount of time, *in the special way that a living organism can*. All self-organizing thermodynamic systems exploit the potentialities of nature in this way, to some extent: otherwise they would not have come into being. But living organisms contain within themselves a special sub-system whose "dedicated," or system-specific, role it is to track the difference between *successful* and *unsuccessful* attempts to exploit the potentialities of nature, and to guide the organism accordingly, towards success and away from failure, including all the degrees from simple devolution or regress, to decay, to death.

In other words, living organisms inherently obey, realize, and satisfy various *normative constraints* on their natural activities. The simplest or most basic normative constraint is the difference between

(i) successfully surviving/living, and
(ii) not surviving/dying.

And under the assumption that normative constraint (i) has been satisfied, then the next most basic normative constraint is the difference between

(ia) successfully reproducing, and
(ib) not reproducing.

And so-on. In turn, non-rational minded animals and rational minded animals alike, whether human or non-human, are nothing more and nothing less than living organisms that are, relatively speaking, *highly attuned to normativity in all its natural dimensions*, hence they have what I will call *finegrained normative attunement*. Indeed, according to the metaphysical thought I am developing here, then this exceptional level of normative attunement in the production of self-organizing thermodynamic activity—our finegrained normative attunement—is precisely, in the lives of rational human minded animals, what *our* consciousness, *our* intentionality, *our* caring, *our* rationality, and *our* free agency *are*.

So, given the notion of a self-organizing thermodynamic system, Non-Local Compatibilism tells us that much and perhaps even most, but *not* all, of physical nature is made up of deterministic natural automata or natural machines, whose causal activities not only satisfy the general causal laws governing the conservation of total-quantity-of-matter-and/or-energy, but also are necessarily determined by and Turing-computable from all the settled quantity-of-matter-and/or-energy-facts about the past, especially including The Big Bang, together with all the general deterministic causal laws. Contrastively and correspondingly, Local Incompatibilism tells us that some, but not all, of physical nature is made up of free agents that are themselves nothing more and nothing less than rational or non-rational human or non-human minded animals, that is, self-organizing thermodynamic systems—in a word, naturally purposive or teleological systems—that are also organisms with finegrained normative attunement, that in turn are *not* deterministic natural automata or machines, even though their activities are all perfectly consistent with the general causal laws of nature, especially including the Conservation Laws, and all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang.

Nor are such minded animals mere *in*deterministic natural automata or natural machines, whose activities not only satisfy the Conservation Laws, but also are necessarily determined by and Turing-computable from all the settled quantity-of-matter-and/or-energy-facts about the past, especially including The Big Bang, together with the general *probabilistic or statistical* causal laws of nature, which make up the rest of physical nature, in addition to deterministic systems and naturally purposive or teleological systems, including minded animal free agents.

In this way, Local Incompatibilism and Non-Local Compatibilism are perfectly logically and really metaphysically consistent with one another, as well as being importantly distinct from classical or strict Compatibilism and classical or strict Incompatibilism alike. Even more excitingly, however:

> (i) an event's consistency with all the deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, captures the truth of the well-known, but not always correctly-understood, principle of *The Causal Closure of the Physical*, aka CCP, which, in its simplest formulation, says that "only physical things can cause physical things,"[47]
> (ii) Non-Local Compatibilism, together with the thesis that there actually exist at least some indeterministic natural automata, captures the truth of *The Scientific Image*, and
> (iii) Local Incompatibilism captures the truth of *The Manifest Image*.

Indeed, consistency with all the deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang—hence CCP—together with Non-Local Compatibilism, together with the thesis that some indeterministic natural automata actually exist, together with Local Incompatibilism, collectively represent the *smooth fusion* of the two Sellarsian world-images, which is something that Sellars himself was never able to do.[48]

If Natural Libertarianism is true, then, The Big Bang—or whatever it was that actually constituted and determined the causal-thermodynamic and nomological origins of the physical universe—together with the physical universe's non-equilibrium, spatiotemporally asymmetric, unidirectional, complex, thermodynamic evolution and cosmological expansion, including entropy,[49] that is, all the settled quantity-of-matter-and/or-energy facts about the past in addition to The Big Bang, together with all the deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, provides the metaphysically necessary causally enabling and constraining background of all the consciously-experienced choices and acts belonging to my unique, complete, and finite self-organizing thermodynamic rational human minded animal life, both now and until I die. But at the same time, I myself am the "ultimate source" of all those choices and acts, and they are all "up to me." Those choices and acts are all *deeply free*. In turn, these deeply free choices and acts do not add any new quantities of matter and/or energy to nature: necessarily, the total quantity of matter and/or energy is always conserved from natural cause to natural effect. Instead, these deeply free choices and acts spontaneously add *new, personally-authored, creative immanent structurings and re-structurings of matter and/or energy* to physical nature, and thereby constitute our own personalized counterpoise to entropy, our own "negentropic sourcehood." Our direct contribution to physical nature is *immanent-structural, not quantitative*.

It is crucial to note, therefore, that the Incompatibilistic Compatibilist metaphysical picture ensures that as a free agent I am not in any sort of *causal competition* with the rest of physical nature, even if nature can, of course, accidentally upset my best laid plans, not to mention those of mice. I am not self-causing or self-creating; I never bring any new quantities of matter and/or energy into nature; and I do not violate any general causal laws of nature, including the stochastic laws. On the contrary, The Big Bang really metaphysically requires *The Little Bang* that is my complete, finite, and unique biological, neurobiological, and rational human minded animal or real human personal life, together with all my life's proper parts, each of them an

even littler little bang, *in order to complete* the existence and specific character of the physical constitution of nature as a whole. This completion is the same as sufficiently augmenting the total immanent structural matter-and/or-energy-grid of physical nature, thereby fleshing out its minimal cosmic conserved-total-quantity-of-matter-and/or-energy skeleton, and reversing entropy. Dare to think and act for yourself! You must change your life! Be authentic and principled! Be autonomously free in the Kantian sense! Reverse entropy with a good will! These are all, ultimately, necessarily equivalent imperatives for rational human animals and real persons like us, deeply free moral agents all.

In this way, each one of us is *a small-scale causal singularity with finegrained normative attunement* via her own free agency, and it is really metaphysically impossible for the universe to be whatever it actually is, without us. For example, the specific character of the fully-elaborated matter-and/or-energy grid of the physical universe right here and now really metaphysically requires that I spontaneously intentionally moved my arm *thus* in such a way that my coffee cup would be precisely *there*, precisely *then*. And ditto, mutatis mutandis, for the local artisan who brought it about that my favorite hand-painted San Francisco coffee cup would be shaped and colored in precisely *these* ways.

For those who think about philosophical theology, it should be obvious how this line of argument is in some ways analogous to the classical *design argument* and the neo-classical *fine-tuning argument* for God's existence.[50] It is clearly fallacious to infer from even massively widespread finite *local design* to the across-all-possible-worlds necessary existence of a single, infinite, omniscient, omnipotent, omnibenevolent *global designer*—hence the classical design argument and neo-classical fine-tuning arguments for God's existence are both unsound. Nevertheless, it is perfectly reasonable to infer the actual-world-bound necessary existence of many finite, non-all-knowing, non-all-powerful, non-all-good *local designers* whose self-organizing and in some cases also freely intentionally-driven limited causal powers are nevertheless precisely sufficient for the production of fully fine-grained local design and matter and/or energy configuration.

Here is what I mean by that. Suppose, for a moment, that we can know a priori that it is really impossible to know or prove either that God exists or that God does not exist.[51] Then, for all we know and for all we don't know, we live in a God-less universe. In such a universe, it is, surely, philosophically unbelievable and even metaphysically really impossible that The Big Bang *causally did it all*, including my writing this sentence. But by sharp contrast, it is also fully philosophically believable and even strongly metaphysically necessary that The Big Bang, *only together with the natural history that has followed it, including each and every one of the many "locally designing" Little Bangs*, causally did it all. Indeed, *I* wrote that sentence, and the one following it, and this one too, not anyone or anything else, and certainly not The Big Bang. What The Big Bang does is to provide the original cosmic matter and/or energy burst, or "initial conditions," and establish the universal causal conservation-of-quantities-of-matter-and/or-energy constraints, including the minimal cosmic skeleton of general causal laws, that are then implemented, over orientable space and irreversible time, in a universe of far-from-equilibrium, complex thermodynamic systems. But subsequent natural history and the many "locally designing" Little Bangs, especially including us and our free agency, *do all the rest of the causal work.*

More precisely, and very importantly, this "local design argument" or "local fine-tuning argument" means that in every relevant metaphysically really possible world, our own "human, all too human" locally designing Little-Bang causal powers would still exist and have the same

specific characters. And if, again contrary to fact, rational human minded animals were to be removed from the actual spacetime of the natural world, in the sense that their complete, finite, and unique conscious, intentional, caring, and rational lives were somehow literally carved out of nature, or somehow rendered causally inert, then it would be just as if The *Big* Bang had been literally carved out of nature, or rendered causally inert, and the actual universe would be strongly metaphysically impossible. Our unique contribution to nature is that necessarily, the physical universe would not have been constituted in just *this* fully fine-grained way, without our free choices and acts. Otherwise put, according to Natural Libertarianism, the complete, finite, and unique lives of deeply free rational human minded animals are metaphysically necessary proper parts of the basic causal-nomological, thermodynamic, and biological architecture of actual physical nature in its fully fine-grained existence and specific character, its complete, densely-packed, asymmetric matter and/or energy-architecture. That is our special *creative co-authorship role* in bringing about the actual natural, physical order of things.

And this insight, in turn, is what finally fully fuses the Scientific and Manifest Images: the actual natural, physical order is completed by all and only those creatures whose nature cannot be *wholly* explained by means of physics, but whose existence and specific character are also completely *consistent* with physics. Physics is the mathematically-driven empirical science of matter-and/or-energy-quantity and matter-and/or-energy-structure. But physics cannot, even in principle, necessarily determine or predict all the precise *structurings* of matter and/or energy in far-from-equilibrium, asymmetric, complex, self-organizing thermodynamic systems. Otherwise put, the physical world achieves causal-thermodynamic and explanatory *completeness* only by means of essentially including what is what is really metaphysically *primitive* and not itself physically reducible to matter-and/or-energy-quantity facts: far-from-equilibrium, asymmetric, complex thermodynamics, especially including self-organization, organismic life, consciousness, intentionality, caring, rationality, deeply free agency, and real persons. In this sense, in an apparent paradox, it seems that physics will finally achieve its ultimate ideal of Grand Unified Theory only by fully incorporating a metaphysically and epistemologically correct conception of *its own explanatory limits*, thereby adopting the attitude of *natural piety*.[52] I will have more to say about this apparent paradox in chapter 2.

It is perhaps needless to say—but I will say it anyway—that the actual natural, physical completion that we, as locally designing Little Bangs and creative co-authors of nature, can provide by means of our deeply free agency in the world, and thereby by means of our own complete, finite, and unique lives, *can be for better or for worse*. We can do it well, with purity of heart, singlemindedness, or wholeheartedly for the sake of our own moral principles and the moral law; or we can do it half-heartedly, double-mindedly, lacking all heart, just badly, or in direct contravention of the moral law—which is to say, more or less inauthentically, simply unsuccessfully, or downright wickedly. No matter how we do it, however, whether for better or for worse, and whether authentically or inauthentically, nevertheless our lives not only have causal efficacy, but also have inherent value and inherent meaning.

You may seriously doubt the truth of this last claim. But now ask yourself just precisely what it is that you are doing when you seriously doubt that your own life has inherent value and inherent meaning, and you begin to think of yourself as, and to feel like, nothing but a biochemical puppet or a moist robot. If I am correct, then only a minded animal with an inherently meaningful life could come to believe deep in its heart, tragically, that its own life is valueless and meaningless. Only a *non*-automaton could ever seriously imagine that it is nothing but a natural automaton and therefore that its own life is absurd—only a *non*-automaton

could ever have a genuine "existential crisis."[53] If all that sounds too telegraphic right now, not to worry: I have much, much more to say about these and other dramatically life-changing and life-informing choices, actions, and emotions in chapter 3 below, in *Kantian Ethics and Human Existence*, and in *Kant, Agnosticism, and Anarchism*.

1.2 Deep Freedom and Principled Authenticity

Here is a third way of putting the very same basic idea behind Natural Libertarianism. Natural Libertarianism is a specifically contemporary *Kantian* theory of free will and practical agency,[54] and it also includes a metaphysical interpretation of a specifically *Kierkegaardian* conception of choice. I will explicate and defend the contemporary Kantian theory of free will and practical agency in the next two chapters. But for the moment I want to focus on the Kierkegaardian part. My Kierkegaardian conception of choice says that choice essentially entails only a strict "Either/Or," even when, in context, this does *not* include "alternative possibilities" in the classical sense of "branching futures" or "open doors":

> Yes, I see it all now perfectly: there are two possible situations—one can do either this or that. My honest opinion and my friendly advice is this: Do it or do not do it—you will regret both…. I say simply Either/Or.[55]

According to the classical branching futures or open doors conception of alternative possibilities, it is a necessary condition of free will and moral responsibility alike, with respect to any of my choices or acts, that I be able to choose or do *X* or *Y*, holding everything else about the world fixed—even if I have no personal interest whatsoever in choosing or doing *X*-or-*Y*, nor any reasons whatsoever for choosing or doing *X*-or-*Y*, and even if, on the contrary, all my personal interest and all my reasons are centered on choosing or doing *X alone*, assuming that I am personally interested in or have any reasons to choose or do anything at all in that context. Prima facie, that seems absurd: why should free will and moral responsibility ever require the existence of *utterly irrelevant alternatives*? So the other conception, the one that I will defend, is the Kierkegaardian Either/Or conception:

> The choice itself is crucial for the content of the personality: through the choice the personality submerges itself in that which is being chosen, and when it does not choose, it withers away in atrophy.
>
> The Either/Or I have advanced is, therefore, in a certain sense absolute, for it is between choosing and not choosing. But since the choice is absolute, the Either/Or is absolute…. I wish only to force you to the point where the necessity of making a choice manifests itself and therefore to consider existence under ethical qualifications.[56]

By sharp contrast to the branching-futures or open-doors conception, then, the Kierkegaardian Either/Or conception says that as long as there is at least *one* sure thing I can choose or do, or not, namely *X*, then I can *either* commit myself to *X*, *or* not. I will call this *X*, the one sure thing I can choose or do, or not, *the live option*. Otherwise put, I am saying that insofar as I have free will, then over and above any abilities whatsoever that I might have at any given time to choose or do *X* or to choose or do *Y*, hence over and above any real alternative

possibilities there may or may not be in that context, there remains another more fundamental capacity I have, which is just the capacity at that time either to exercise my ability to choose or do the live option *X*, or not exercise my ability to choose or do the live option *X*, and this is my capacity for *self-commitment*, or for what Kierkegaard calls "choosing myself."[57]

According to the Kierkegaardian Either/Or conception, for the purposes of free will and moral responsibility it simply does not matter either metaphysically or epistemically whether, in that context, there is any alternative *Y* to my choosing or doing *X*. This is because sometimes, as a matter of fact, in that context, there really is only one thing, *X*, the live option, that I *can* choose or do, or not, because any other physically possible branching-future or open-door alternatives have been temporarily shut down or closed. To use Locke's famous example, it might be that I am unknowingly inside a room whose door is locked from the outside—nevertheless it may also be that I never seriously consider anything but staying inside that room. So then I freely choose to stay in the room, or not, as the case may be.

And sometimes, even when, in context, there really are several things I could possibly choose or do, or not, in fact there is only one thing I would ever really want to choose or do, or not, or ever really have any reason to choose or do, or not, namely *X*, the live option for me in that context. For example, it is just an actual historical, psychological, honest-to-goodness fact about me that I never seriously considered any career other than talking, teaching, and writing philosophy, once I encountered real philosophy in late high school and college. I fell in love with it like I had been hit by a ton of bricks, complete with shakes, shivers, and sleepless nights thinking about the right and the good, beauty, truth, knowledge, logic, the mind-body problem, freedom and determinism, life-changing metaphysics more generally, Plato's *Dialogues*, Kant's first *Critique*, or Wittgenstein's *Tractatus*. Philosophy was the only thing I ever really wanted to do "for a living." It was the only live option for me: alternative possibilities were completely irrelevant to me in that context. It was only a question of whether I would really and truly choose and do philosophy—or not, game over. So all I ever wanted—as a full-time, lifetime, calling—was *one door* that I could open, or refrain from opening.[58]

Now on the supposition that either of these sorts of situation actually obtains in some context—that is,

> either (i) the Lockean No-Real-Alternatives sort of situation in which staying inside the locked room, with no real alternatives, is all someone ever really wants, or (ii) the Single-Minded-Option sort of situation in which opening one door, the one single-minded option, is all that someone ever really wants or has any reasons to choose or do, or not,

—and if I am that person, then *X* is my *live option* and if and only if

> (1) I can commit myself to choosing or doing X, or not, (or: I could have committed myself to choosing or doing X, or not),
> (2) X would never actually happen unless I were to choose it or do it, (or: X would never have actually happened unless I had chosen it or done it), and
> (3) I actually choose and do X, or not (or: I actually chose and did X, or not),

from which it directly follows that I am (or: was) still fully free and fully morally responsible for choosing or doing *X*, or not. So, in particular, if *X* happens (or: happened), then it flows (or: flowed) directly and ultimately from my self-commitment to it.

My idea is that this metaphysically-interpreted Kierkegaardian conception, when taken together with the contemporary Kantian conception of anti-mechanism, free will, and practical agency, that I will spell out in chapters 2 to 5 below, jointly yield what I call *deep* freedom of the will. Otherwise put, if I am deeply free, then I am the *ultimate source* of all my choices and intentional acts, and those choices and acts are all *up to me*, for better or worse. If whatever I choose and do were ultimately caused either by The Big Bang or by a more local environmental process or state of the physical world, then I would be distally or proximally determined and not deeply free. That, again, is Local Incompatibilism and its incorporation of The Manifest Image. But at the same time, I can freely choose and do whatever I freely choose and do, only as presupposing and as necessarily enabled by the fact there is a fairly massive background of deterministic general causal laws—especially including the Conservation Laws—and deterministic processes that I can exploit for my own natural purposes, together with some indeterministic natural automata, out there in the natural world, always assuming, of course, that the local natural conditions are propitious, such that I have sufficient non-deterministic "natural open space" for my "live options," and that I have a far-from-equilibrium, spatially orientable, temporally irreversible, complex, self-organizing, organismic thermodynamic life that necessarily includes my consciousness, intentionality, caring, and rationality as its specific dominant immanent structures. That, again, is Non-Local Compatibilism, and its incorporation of The Scientific Image.

One last very important point in this connection. If Natural Libertarianism is correct, then because our deep freedom is freedom-in-life, it follows that as long as you are consciously, intentionally, caringly, and rationally alive, then the scope of your free agency is always a matter of varying amounts of natural open space, more or less, and is never merely binary or "on-off"—like a glorified light switch. In my opinion, you are *not* just a glorified light switch. You are *not* just a Turing machine, even a fleshy one tricked out with all sorts of epiphenomenal "reasons responsive" bells and whistles. Your life is *not* necessarily determined in all its natural properties and causal powers by means of a closed set of Turing-computable algorithms expressing all the deterministic or stochastic general causal laws of nature, especially including the Conservation Laws, together with all settled matter and/or energy-facts about the past, especially including The Big Bang. In addition to all your computational abilities and processes, in addition to all your "reasons responsive" mechanisms, and most fundamentally, you are a far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic thermodynamic system with fine-grained normative attunement, and a minded animal, and therefore you are capable of and physically realize various spontaneously generated, non-mechanical, and uncomputable—and in particular, naturally purposive or teleological—processes. Above all, you are a *rational human* minded animal and a *real human person* with always at least one "live option" available to you—Either/Or. This is true even when the amount of "natural open space" that is currently actually available to you, by dint of brute factual circumstances and bad luck, has been reduced almost (but not quite) to zero, and when what A.N. Whitehead so aptly called "the goading urgencies of contingent happenings" have been, as it were, merrily rammed down your throat. As long as you are consciously, intentionally, caringly, and rationally alive, as long as you can still breathe, feel, desire, emote, and think, then there is always something you can choose and do, or refrain from so choosing

or doing, even if it is only just choosing your affirmation or denial of whatever the rest of the world is merrily ramming down your throat at that moment. As long as you are consciously, intentionally, caringly, and rationally alive, you can always have purity of heart. You can always be single-minded. You can always be wholehearted. You can always choose yourself. You can always change your life. You can always be *authentic*.

By a categorical contrast, *inauthenticity* is comporting yourself *as if you were a natural automaton*—as if you were really and truly a biochemical puppet or moist robot, a fleshy deterministic or indeterministic Turing machine running a complicated decision-theoretic program, relentlessly causally conserving total quantities of matter and/or energy, relentlessly furthering the causal impact of The Big Bang, and *not* really and truly alive; as if you were *not* a real person; as if you could *never* think or choose or act for yourself; and as if you did *not* really have the capacity for deep freedom and for achieving or realizing principled authenticity, at least partially and to some salient degree or extent.

That is the deep metaphysical and moral insight offered by Kierkegaard. Later, we will see that essentially the same deep morally-driven insight about the metaphysical relationships between natural science (including physics, chemistry, biology, and neurobiology) and the natural universe, free agency, real personhood, morality, authenticity, and inauthenticity is also developed in slightly different ways by Nietzsche, Wittgenstein, Camus, Sartre, and Harry Frankfurt. But above all, this deep insight is developed by Kant, at least as I understand him, in full view of the classical 19th, 20th, and 21st century Analytic and existential-phenomenological (aka "Continental") traditions in philosophy that have flowed as bifurcating (and now, seemingly, never-to-be-reunited) branches from the core Kantian and post-Kantian (including German idealist and neo-Kantian) European philosophical tradition. A consensus in belief among several great thinkers, no matter how great they might be, does not of course entail the truth of that belief. Indeed, shared belief, in and of itself, does not entail truth of *any* sort, except, of course, the truth that there are some shared beliefs—no matter what cultural relativists and social constructivists may say. Nevertheless, I do think that in this case, collectively, Kant, Kierkegaard, Nietzsche, Wittgenstein, Camus, Sartre, and Frankfurt, as I understand them, are all onto some things of fundamental importance about the nature of our rational human animal lives. Those shared thoughts will make a number of special guest appearances as this book goes on. And in *Kantian Ethics and Human Existence* and in *Kant, Agnosticism, and Anarchism*, I also spell out in detail what I take to be their ultimate moral, theological, and political pay-offs.

1.3 The Central Claim of this Book, and Previews

The central claim of this book is that the full metaphysical and normative power of a Kantian theory of human free will, practical agency, and persons, with deep and radical ethical and political implications downstream, can be captured by a correct understanding of how biological life in general, and the lives of human minded animals in particular, relate to the rest of physical nature. Or otherwise put, the central claim of this book is that Natural Libertarianism and Minded Animalism are both true.

Now that the basic ideas behind Natural Libertarianism and Minded Animalism are on the table, I need to begin to re-describe them more carefully, and to argue for them step-by-step.

By way of previews, here is an outline of my overall six-step argument for Natural Libertarianism, with each step labeled in italics and keyed to chapter titles for convenient back-reference, followed by an annotated outline of the other chapters beyond this Introduction, which doubles as chapter 1.

The Six-Step Argument for Natural Libertarianism

(1) Beyond Mechanism.

Biological life is a physically irreducible but also non-dualist and non-supervenient necessary a priori immanent structure of a well-defined class of far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic systems. (Premise, justified in chapter 2.)

(2) From Biology to Agency.

Free rational minded animal agents are nothing more and nothing less than conscious, intentional, caring, rational self-organizing, organismic thermodynamic systems that are capable of (i) deeply free choice based on effective desires and instrumental or non-instrumental internal reasons,
(ii) autonomy in the Kantian sense, or rational self-legislation, and
(iii) authenticity in the Existentialist sense, namely, purity of heart, single-mindedness, or wholeheartedness. (Premise, justified in chapter 3.)

(3) Neither/Nor.

Natural Mechanism is the weak disjunctive combination of Universal Natural Determinism and Universal Natural Indeterminism. More specifically, something is naturally mechanized, or a natural automaton, if and only if all its causal behaviors, functions, and operations are necessarily determined by all the deterministic or probabilistic/statistical general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and-or-energy facts about the past, especially including The Big Bang, and Turing-computable from that base. But not everything natural is Conservation-Laws-determined, Big Bang-caused, and Turing-computable. So Natural Mechanism is false, hence both Universal Natural Determinism and Universal Natural Indeterminism are false. Morever, Hard Determinism is false. Soft Determinism is false. And Classical Libertarianism (including its agent-causal, non-causal, and event-causal versions) is false. Correspondingly, classical Compatibilism (including Soft Determinism, Semi-Compatibilism, Revisionism, and In-the-Zone Compatibilism) and classical Incompatibilism (including Hard Determinism, Hard Incompatibilism, Classical Agent-Causal Libertarianism, Classical Non-Causal Indeterminist Libertarianism, and Classical Event-Causal Indeterminist Libertarianism) are all false. At the same time, Local Incompatibilism and Non-Local Compatibilism are true. So Incompatibilistic Compatibilism is true. (Premise, justified in chapter 4.)

(4) Either/Or.

Harry Frankfurt's argument against The Principle of Alternative Possibilities (PAP) is sound, but it does not follow that deep (non-)moral responsibility is compatible with either Universal Natural Determinism or Universal Natural Indeterminism, since Natural Mechanism is false. On the contrary, the Kierkegaardian Either/Or, which flows from the capacity for self-commitment to a live option, is presupposed by all Frankfurt-style counterexamples to PAP; and as metaphysically embedded in a larger free-agency-structure which also includes the capacities for veridical psychological freedom and for principled authenticity (namely, the capacity for autonomy in the Kantian sense, or rational self-legislation, together with the capacity for purity of heart, single-mindedness, or wholeheartedness), this capacity for self-commitment to a live option, or Kierkegaardian Either/Or, is a necessary and sufficient condition of deep (non-) moral responsibility. (Premise, justified in chapter 5.)

(5) Deep Freedom and Principled Authenticity.

The capacity for self-commitment to a live option, or the Kierkegaardian Either/Or, along with the capacities for veridical psychological freedom, real causal spontaneity, and ownership, are necessary and sufficient conditions of the capacity for deep freedom. In turn, the capacity for deep freedom is a necessary but not sufficient condition of the capacity for principled authenticity, which, as incorporating deep freedom, yields deep (non-)moral responsibility. (Premise, justified in chapter 5.)

(6) Natural Libertarianism.

Therefore, since Natural Libertarianism is just the three-part thesis (i) that freedom is in life, (ii) that Incompatibilistic Compatibilism is true, and (iii) that the constitution of free rational human minded animal agency inherently includes the capacities for deep freedom and principled authenticity, together yielding deep (non-)moral responsibility, then it follows that Natural Libertarianism is true. (Conclusion, from premises 1-5 above.)

An Annotated Outline of the Other Chapters

Chapter 2. Beyond Mechanism: The Dynamics of Life

This chapter works out an anti-mechanistic, non-physicalist, non-dualistic approach to the philosophy of biology—which I call *dynamicism*[59]—by deploying some Kantian ideas from the *Critique of the Power of Judgment*, in particular, the idea that organismic life is inherently goal-directed, naturally purposive, or naturally teleological (aka "self-organizing"), but also fully situating these ideas in the framework of contemporary theories of the nature of life. The primary purpose of this chapter is to show that free agency, as a capacity of minded animals, requires uncomputable, spontaneous, causally efficacious, far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing,

egocentrically-centered, reproductive, growing and decaying, more-or-less motile, metabolizing, evolutionary or naturally-selecting, epigenetic, more-or-less finegrainedly normatively attuned thermodynamic activity as a necessary immanent structural ground.

Chapter 3. From Biology to Agency

This chapter uses the non-reductive, Kant-inflected, dynamicist philosophy of biology developed in chapter 2, and combines it with a similarly Kant-inflected theory of intentional agency and practical freedom, but also combines it with Frankfurt's hierarchical-desire theory of the will as a way of elaborating a broadly Kantian theory of practical reasoning in a contemporary context. The result is a fully-rounded theory of free agency that is at once naturalistic, non-reductive, and psychologically realistic.

Chapter 4. Neither/Nor: The Negative Case for Natural Libertarianism

This chapter provides a critical survey of some important theories and debates in contemporary work on free will, including mainstream versions of Hard Determinism, Soft Determinism, and Classical Libertarianism (including its agent-causal, non-causal, and event-causal versions), and also standard arguments for and against Compatibilism or Incompatibilism. Along the way, a crucial distinction is made between (i) *determinism* (that is, all the settled quantity-of-matter-and/or-energy facts about the past together with the general causal laws of nature necessitate a single closed present and future), (ii) *indeterminism* (that is, all the settled quantity-of-matter-and/or-energyfacts about the past together with the general causal laws of nature do not necessitate a single closed present and future but nevertheless still necessitate a fixed probability space of open presents and futures, according to statistical causal laws of nature), and (iii) *non*-determinism plus *non*-indeterminism (that is, there are some natural processes that are not deterministic, but also not *in*deterministic, and not *mechanistic*, but instead inherently involve the ability of a self-organizing thermodynamic system, for example, a living organism, to exploit some open texture in nature—aka "natural open space"—and create something relatively original and unprecedented in the natural world, without violating any general causal laws of nature, especially including the Conservation Laws). The positive result of the chapter is to carve out a place in logical space for Natural Libertarianism as a version of non-determinism that is incompatibilist with respect to minded animal agents and their lives (especially rational human animals or real human persons, but also non-rational human or non-human minded animals), but also compatibilist with respect to much or most of physical nature apart from living organisms and minded animals, hence as Incompatibilistic Compatibilism.

Chapter 5. Either/Or: Deep Freedom and Principled Authenticity

This chapter spells out and argues for Natural Libertarianism in detail, with special reference to the nature of non-deterministic, non-indeterministic, locally incompatibilistic deep freedom, "up-to-me-ness," or "ultimate sourcehood." The

theory of deep freedom also incorporates a Kierkegaardian "Either/Or" theory of causally and morally responsible choice. This kind of choice does not require alternative possibilities, but instead requires only what I call a *live option*. By this, I mean that in context there is one thing *X* that I can either commit myself to choosing or doing, or not commit myself to choosing or doing, and *X* would not happen (or: would not have happened) if I did not (or: had not) thereby commit(ted) myself. Therefore, as long as I have a live option—as it were, one door that I can open, or refrain from opening—then my choice or act is deeply free, even if, in that context, there are no alternative possibilities. This chapter fuses the Kantian/Kierkegaardian theory of deep freedom with the Kantian/Frankfurtian theory of practical agency developed in chapter 3. It also adds an Existentialist-inspired theory of authenticity vs. inauthenticity to the Kantian/Frankfurtian theory of practical agency, in order to capture the full depth and texture of rational human free agency. A crucial feature of the overall account is that only real persons with sufficiently unified lives can count as deeply free subjects of causal responsibility and deep (non-)moral responsibility, and correspondingly that only real persons with sufficiently unified lives can count as principled, authentic rational free agents. That in turn provides a natural segue to the metaphysics of persons and personal identity worked out in chapters 6 to 7.

Chapter 6. Minded Animalism I: What Real Persons Really Are

The argument of this chapter builds upon three interconnected propositions: the theory of real human persons called *Minded Animalism*; a corresponding proposal for a set of necessary and sufficient conditions for real personhood; and a corresponding proposal for a set of criteria for real personal identity over time (diachronic identity) and at a time (synchronic identity). The four core ideas behind Minded Animalism are (i) that real human persons are conscious, essentially embodied living organisms within the human species, that also possess capacities for intentionality, caring, self-consciousness, and rationality, (ii) that there are two distinct types of real persons (including, on the one hand (a) all normal human infants and young children, some impaired older children and impaired human adults, and certain species of non-human animals, and on the other hand (b) all normal older children and most human adults), depending on their differently configured and disposed, or impaired/unimpaired, online capacities for instrumental and/or non-instrumental rationality, (iii) that the identity relation for real persons is an identity relation between proper parts or wholes of minded animal lives, not between minds and minds, or bodies and bodies, or even animals and animals, and (iv) that a necessary condition of real personhood and real personal identity alike is that a real person's minded animal life be constituted, at least in part, by sufficiently many deeply free choices and acts. This last feature directly and explicitly connects the metaphysics of free will with the metaphysics of persons and personal identity.

Chapter 7. Minded Animalism II: From Parfit to Real Personal Identity

This final chapter critically compares and contrasts Minded Animalism with Derek Parfit's highly influential theory of persons and personal identity in

Reasons and Persons, as a theoretical foil for bringing out various important elements and philosophical virtues of the Minded Animalist view. This chapter also re-emphasizes the fundamental connections between Natural Libertarianism and Minded Animalism, which are that, insofar as the basic conditions on our Kantian/Kierkegaardian deep freedom and our Kantian/Frankfurtian practical agency are satisfied, so too the basic conditions on our identity as real persons are satisfied, and conversely.

Chapter 2

BEYOND MECHANISM: THE DYNAMICS OF LIFE

It is quite certain that we can never adequately come to know the organized beings and their internal possibility in accordance with merely mechanical principles of nature, let alone explain them; and this is so certain that we can boldly say that it would be absurd for humans to make an attempt or to hope that there could ever arise a Newton who could make comprehensible even the generation of a blade of grass according to natural laws that no intention has ordered; rather we must absolutely deny this insight to human beings. (CPJ 5: 400)

For a phenomenon such as life,… the physical facts imply that certain functions will be performed, and the performance of these functions is all we need in order to explain life…. A vitalist might have claimed that it is logically possible that a physical replica of me might not be alive, in order to establish that life cannot be reductively explained. And a vitalist might have argued that life is a further fact, not explained by any account of the physical facts. But the vitalist would have been wrong… Vitalism was mostly driven by doubt about whether physical mechanisms could perform all the complex functions associated with life: adaptive behavior, reproduction, and the like. At the time, very little was known about the enormous sophistication of biological mechanisms, so this sort of doubt was quite natural. But implicit in these very doubts is the conceptual point that when it comes to explaining life, it is the performance of various functions that needs to be explained. Indeed, it is notable that as physical explanations of the relevant functions gradually appeared, vitalist doubts mostly melted away…. Presented with a full physical account showing how physical processes perform the relevant functions, a reasonable vitalist would concede that life has been explained. There is not even conceptual room for the performance of these functions without life.[60]

If there is anything in the approach I adopt, it will follow that concepts like life, life-form, … etc., have something like the status Kant assigned to "pure" or a priori concepts…. [E]ven if our concept life-form arises with experience, it need not be thought to arise from it; its content is rather supplied by reflection on certain possibilities of thought or predication.[61]

We are in a world of multiple fluctuations, some of which have evolved, while others have regressed. This is in complete accord with the results of far-from-equilibrium thermodynamics…. But we we can now go even farther. These fluctuations are the macroscopic manifestations of fundamental properties of fluctuations arising on the microscopic level of unstable dynamical systems…. Irreversibility, and therefore the flow of time, starts at the dynamical level. It is amplified at the macoscopic level, then at the

> level of life, and finally at the level of human activity. What drove these transitions from one level to the next remains largely unknown, but at least we have achieved a noncontradictory description of nature rooted in dynamical instability. The descriptions of nature as presented by biology and physics now begin to converge.[62]

2.0 Introduction

What is the nature of biological life, and how do we represent it? In this chapter, using Kant's theory of mental representation and his philosophy of biology as starting points, in relation to contemporary theories of mental representation and contemporary philosophy of biology, I am going to argue that there is not only a non-trivial "explanatory gap" but also a correspondingly non-trivial "ontological gap" between reductive or non-reductive physicalist—what I will call "naturally mechanistic"—approaches to biology on the one hand, and the non-equilibrium thermodynamic phenomenon of life on the other.

Now explanatory irreducibility is the irreducibility of the contents of some mental representations to the contents of some other mental representations. Correspondingly but also by contrast, ontological irreducibility is the irreducibility of some worldly properties and/or facts to some other worldly properties and/or facts. Provided that there is a necessary one-to-one connection between distinct mental representations and distinct worldly properties and/or facts, then explanatory irreducibility entails ontological irreducibility. I am going to argue that a contemporary Kantian theory of the mental representation of life, when taken together with a neo-Aristotelian and contemporary Kantian theory of worldly properties, does indeed yield such a necessary one-to-one connection between the distinct representations of living organisms and natural mechanisms on the one hand, and the distinct worldly constitutive properties corresponding to living organisms and natural mechanisms on the other hand, and also that the relevant worldly properties constituting living organisms, as such, are causally efficacious.

As a consequence, then just as the many well-known non-reductive arguments about consciousness that surfaced in the late 20th century forced us seriously to reconsider and rethink our basic commitments and basic notions in the philosophy of mind, so too we must now seriously reconsider and rethink our basic commitments and basic notions in the philosophy of biology. Or otherwise put: having taken *the phenomenon of consciousness* seriously, a fortiori, we must now also take *the phenomenon of life* equally seriously.

At the same time, however, the positive version of anti-mechanism and non-reductionism about biological life that I am going to propose—which I call *dynamicism*—does not involve any epistemological or metaphysical equivalent of Cartesian dualism, whether an ontological dualism (of essentially distinct substances) or a property dualism (of essentially distinct properties and/or facts). This, in turn, rules out not only *reductive* physicalism but also *non-reductive* physicalism, which is committed to a version of property-dualism, and strong supervenience on fundamental physical facts, just as much as it rules out Cartesian substance-dualism. According to the dynamicist model of life, biological life is a *non-equilibrium thermodynamic phenomenon*, and the non-equilibrium thermodynamic phenomenon of life is neither explanatorily nor ontologically reducible to the causal natural mechanisms bound up with fundamental physical properties and/or facts. But at the same time it remains true that the non-equilibrium thermodynamic phenomenon of life is *not essentially distinct from physical*

causal processes. According to the dynamicist view, the phenomenon of life is an inherently non-mechanical, irreducible, necessary a priori *immanent structure* of a well-defined class of far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic thermodynamic physical processes. As such, it is no more distinct from those very physical processes than either their underlying asymmetric *spatiotemporal* causally relevant structure, or their underlying uncomputable *mathematical* causally relevant structure, is distinct from them. Indeed, life is nothing less and nothing more than an ordered sequence of self-organizing, inherently context-dependent, reproducing, growing and decaying, more-or-less motile, evolving or naturally selecting, spontaneous, metabolizing, epigenetic, more-or-less finegrainedly normatively attuned, immanent structurings and re-structurings of flows of matter and/or energy. Hence the non-equilibrium thermodynamic phenomenon of life, as such, is an immanent structure all of whose realizations are not merely causally *relevant* but also causally *efficacious*.

2.1 IMMANENT STRUCTURALISM

The basic idea behind my neo-Aristotelian and contemporary Kantian "immanent structuralist" metaphysics of properties is fairly simple—

> immanent structuralism: Some (kinds of) material objects, events, processes, and facts have causally relevant or causally efficacious structural properties that are not strictly determined
>
> either by (i) the intrinsic non-relational properties of their physical parts together with all the extrinsic non-relational or relational properties of their physical parts, or by (ii) the total set of actual contingent sensory-experiential or natural objects, events, and facts.

The properties satisfying these conditions are what I call *intrinsic structural properties* or *immanent structural properties*, and they include—

> (i) all fundamental mathematical properties of material objects, events, processes, and facts,
> (ii) all fundamental asymmetric spatiotemporal and non-equilibrium, complex, self-organizing thermodynamic properties of material objects, events, processes, and facts, and energy, especially including fundamental biological properties of living organisms, and
> (iii) all fundamental mental properties of animals, including consciousness-properties, intentionality-properties, caring-properties, rationality-properties, and free-agency-properties.

When immanent structural mental properties of human animals are appropriately combined with human linguistic facts and social facts, then they (namely, the appropriate combinations of immanent structural mental properties, human linguistic facts, and human social facts) are essentially the same as what John Searle calls *constitutive rules*. Constitutive rules are rules, as Searle puts it, that

> create or define new forms of behavior. The rules of football or chess, for example, do not merely regulate playing football or chess, but as it were create the very possibility of playing such games.[63]

Searle's view is about linguistic facts and social facts, and more generally about intentional action. But my view is about *natural* facts, and not merely about linguistic and social facts. More precisely, on my view, this constitutive-rule function is present in all immanent structural properties of material objects, material events, material processes, material facts and/or energy facts. Immanent structural properties "create or define" new forms of non-equilibrium, spatiotemporally asymmetric, complex thermodynamic existence or movement, and thereby "create the very possibility of" such causal-dynamic patterns. When instantiated in actual space and time, immanent structural properties *actualize* these new forms of non-equilibrium, asymmetric, complex thermodynamic existence or movement. Thus immanent structural properties, whenever and wherever they exist either in an uninstantiated mode or else via their instances in the natural world, are *neither* "downwardly identical" to any of the other physical properties of some material objects, facts, or events, *nor* "strongly supervenient" on them, *nor* anything otherwise "over and above" them, since they are just irreducible and synthetic a priori necessary structures *of* those physical things and *inherently in* those physical things.

Otherwise put, my claim is that fundamental natural structures of physical things and thermodynamic processes are at once fully irreducible to and also fully immanent in the physical things and thermodynamic processes that are "created or defined by," namely, *constituted by*, those very structures, and *thereby* those very structures are either causally efficacious or at least causally relevant. That is the basic idea. Maiese and I have already argued for immanent structuralism in the philosophy of mind in *Embodied Minds in Action*, by means of demonstrating the existence of immanent structural causally relevant, causally efficacious *mental* properties in suitably complex living animal bodies—a view we call "neo-Aristotelian hylomorphism." Correspondingly, I have also argued for immanent structuralism in the philosophy of logic and mathematics in *Cognition, Content, and the A Priori*, by means of arguing for the existence of immanent structural causally relevant (although not causally efficacious) *logical* and *arithmetic* properties—a view I call "Kantian structuralism." Following on from those lines of reasoning, the purpose of the present chapter is to argue for immanent structuralism in the philosophy of biology, by means of arguing for the existence of immanent structural causally relevant, causally efficacious *organismic* properties.

More precisely, if I am correct, then irreducible organismic life is nothing more and nothing less than an inherently non-mechanical, constitutive-rule-like, immanent structural property of the causal behaviors, functions, and operations bound up with fundamental physical properties and facts in far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic systems of a well-defined class, corresponding to an a priori formal representation of life.[64] In turn, this happens in just the way that, according to Aristotle's hylomorphism, actualizing form relates to the potentiality of the specific matter that it constitutively informs; and it also happens in just the way that, according to Kant in the Transcendental Aesthetic section of the *Critique of Pure Reason*, space and time are nothing but necessary a priori constitutive immanent structural properties of the causally efficacious objects of human experience, corresponding to pure subjective forms of sensible intuition. Hence organismic life is irreducible to natural mechanisms because it is a transcendental,

inherently non-mechanical, naturally purposive or naturally teleological fact about the causal processes bound up with fundamental physical properties and facts in certain far-from-equilibrium, asymmetric, complex, self-organizing thermodynamic systems. But this is *not* because it is an essentially different fact that is something "over and above" the fundamental physical world, and also *not* because it is nothing but a multiply realizable second-order physical fact that is *strongly supervenient* on first-order, fundamental physical facts.

I will stop briefly here to define the notion of "strong supervenience," since it is important in what follows. Strong supervenience[65] is a necessary determination-relation between sets of properties of different ontological "levels," a relation that is weaker than strict property-identity, and is usually taken to be asymmetric, although two-way or bilateral supervenience is also possible. But assuming for the purposes of simpler exposition that supervenience is asymmetric, then, more precisely, *B*-properties (= the higher level properties) strongly supervene on *A*-properties (= the lower-level properties) if and only if

> (i) for any property *F* among the *A*-properties had by something *X*, *F* necessitates *X*'s also having property *G* among the *B*-properties (upwards necessitation), and
> (ii) there cannot be a change in any of *X*'s *B*-properties without a corresponding change in *X*'s *A*-properties (necessary co-variation).

It follows from strong supervenience that any two things *X* and *Y* share all their *A*-properties in common only if they share all their *B*-properties in common (indiscriminability). In turn, *logical* supervenience is a super-strong version of strong supervenience which says that the necessitation relations between the *B*-properties and the *A*-properties are *logical and a priori*. Or more simply put: The *B*-properties are "nothing more than" and "nothing over and above" the *A*-properties. If there were such a being as an all-powerful and all-knowing creator God, and if S/He were to create and/or know all the *A*-properties, then S/He would have nothing more to do in order to create and/or know all the *B*-properties.

It is important to recognize that strong supervenience specifies, at most, a set of *extrinsic modal properties and relations* (namely, upwards necessitation, necessary co-variation, and indiscriminability) between a thing's *A*-properties and its *B*-properties, or between any two things' *A*-properties and *B*-properties. If properties or relations of strong supervenience hold for a thing or things, there is no further implication that these are properties or relations of *constitution* or *essence*, such that a thing's or things' immanent structural characteristics—and in particular, if the thing or things are natural or physical, their efficacious causal powers—depend on these properties or relations. Conversely, if properties or relations of constitution or essence hold for a thing or things, then there is no further implication that strong supervenience holds for them. In short, the metaphysics of strong supervenience is modally *shallow*, not modally deep, unlike the *real* metaphysics of constitution or essence.[66]

2.2 Natural Mechanism, Computability, and Anti-Mechanism

The doctrine of what I will call *physicalism about life* says that biological life is at the very least strongly supervenient on (= non-reductive physicalism), and might also be identical with

or logically supervenient on (= reductive physicalism), the inherently mechanical, causally efficacious behaviors, functions, and operations bound up with fundamental physical properties and facts.[67] When this non-reductive or reductive physicalist doctrine is generalized beyond life to every natural phenomenon whatsoever, then it constitutes what I will call the doctrine of *Natural Mechanism*.

But what, more precisely, is *the very idea* of Natural Mechanism? My claim is that there is a deep and indeed essential connection between natural mechanisms, the conservation of total quantities of matter and/or energy from physical causes to physical effects, effectively decidable procedures, recursive functions, and Turing-computability. More precisely, what I am proposing is that anything's causally efficacious behaviors, functions, operations, and/or states are inherently mechanical in both their existence and their specific character if and only if

> (i) they are necessarily determined by all the general deterministic or indeterministic causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, and
> (ii) they strictly conform to The Church-Turing Thesis (aka "Church's Thesis").

Otherwise put, natural mechanism is the conjunction of "*closed" causal-nomological activity and/or states* (that is, causally-nomologically closed with respect to conserved quantities of matter and/or energy) and *computable activity and/or states*.

So formulated, my analysis of natural mechanism is specifically intended to be comprehensive over, and in equipoise with respect to, the now-standard distinction in the vast mechanistic explanation literature, between "etiological" (that is, causal) mechanistic explanations and "constitutive" (that is, ontological) mechanistic explanations.[68]

And what is The Church-Turing Thesis, aka Church's Thesis? To state it clearly but also non-technically, I must define some terms. An *effectively decidable procedure* is a rule-governed, step-by-step process that yields a pre-established determinate result of a binary kind (for example, either 0 or 1) in a finite or countably infinite number of steps. Otherwise put, an effectively decidable procedure is an *algorithm*. This appears to be the very same notion as that of a *recursive function*,[69] and it also appears to be necessarily equivalent with the notion of a *Turing machine*.[70]

It is important to note that strictly speaking, platonically abstract (that is, non-spatiotemporal, causally inert) or purely mental (angelic, ghostly, spiritual, etc.) Turing machines are at least barely conceivable or logically possible. For example, the latter is what Leibniz called an *automaton spirituale*. But in this chapter and in this book I am focusing exclusively on spatiotemporally located, causally efficacious, embodied/incarnate universal Turing-machines, or what I will call *real-world Turing-machines*. Then The Church-Turing Thesis says that every effectively decidable procedure is a recursive function and also a Turing-computable function, which in turn restricts effectively decidable procedures to digital machine computation,[71] given two plausible further assumptions to the effect that

> (i) the causal powers of any real-world Turing machine are held fixed under our general causal laws of nature, especially including the Conservation Laws, and

> (ii) the "digits" over which the real-world Turing machine computes constitute a complete denumerable set of spatiotemporally discrete physical objects.

Therefore, according to my proposal,

> Anything X is a *natural automaton*, or *natural machine*, if and only if
>
> (1) X is constituted by an ordered set of causally-efficacious behaviors, functions, operations, and/or states (aka "causal powers"),
> (2) the causal powers of X are necessarily determined by all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, together with all the general deterministic or indeterministic causal laws of nature, especially including the Conservation Laws, and
> (3) X's causal powers are all inherently effectively decidable, recursive, or Turing-computable, given two further plausible assumptions to the effect that
> (3i) the causal powers of any real-world Turing machine are held fixed under our general causal laws of nature, and
> (3ii) the "digits" over which the real-world Turing machine computes constitute a complete set of mathematically denumerable (that is, non-real-number, non-complex-number, non-transfinite) quantities, that is, spatiotemporally discrete, physical objects.

In an illuminating paper, Arnon Levy plausibly argues that the very idea of mechanism, or "machine-likeness," essentially includes the two elements of what he calls *decompositional explanation* and *causal orderliness*:

> The guiding idea is that machine-like systems are especially amenable to decompositional explanation, i.e., to analyses that tease apart underlying components and attend to their structural features and interrelations. I argue that for decomposition to succeed a system must exhibit causal orderliness, which I explicate in terms of differentiation among parts and the significance of local relations.[72]

In my formulation of the definition of a natural mechanism, the denumerable-discrete-digits and Conservation-Laws-determined causal-powers elements capture, respectively, Levy's notions of "decompositionality" and "causal orderliness," but within the more physically and formally precise contexts of Turing-computability and contemporary physics. It follows, then, that non-mechanical processes are neither "decompositional" nor "causally orderly," in just the ways required for real-world Turing-computability or Conservation-Law-determination. But it does *not* follow from this, that uncomputable, non-Conservation-Law-determined, hence non-mechanical processes do *not* have either rich physical structure or causal efficacy. On the contrary, as we will see, the uncomputable, non-Conservation-Law-determined, non-Big-Bang-caused, hence non-mechanical processes belonging to the lives of living organisms, especially including the free agency of rational minded animals, *can and do* still fully possess both rich physical structure and causal efficacy.

It is also important to recognize that although all specifically *deterministic* natural processes are real-world Turing-computable, *not* all real-world Turing-computable processes are deterministic. Indeed, there are *indeterministic* real-world Turing-machines. More

generally, if an indeterministic natural process implements a step-by-step probabilistic or statistical rule—that is, if the process is *stochastic*—then it is real-world Turing-computable. Therefore, although all naturally mechanistic, Conservation-Law-determined, Big-Bang-caused processes are real-world Turing-computable, nevertheless naturally mechanistic processes can be either deterministic or indeterministic. This, in turn, is the same as to say that each and every one of the causal behaviors, functions, operations, and/or states (namely, causal powers) of naturally mechanistic physical processes is entailed or necessitated by the general deterministic or indeterministic causal laws of nature, especially including the Conservation Laws, together with the set of settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, and is Turing-computable from that "causally-nomologically closed" physical base.

In section 1.1 above, I formulated a fundamental and fully general Kantian distinction between

> (i) an activity's being merely in conformity with (namely, being merely consistent with, acting merely according to) a law or rule, and
> (ii) an activity's being strictly governed by (namely, being strictly entailed or necessitated by, acting strictly from or for the sake of) a law or rule.

This fundamental distinction applies directly to digital or Turing machine computation. More specifically, there is a correspondingly fundamental distinction between

> (i) what is merely correctly describable or can be simulated in Turing-computable terms, and
> (ii) what strictly encodes or inherently implements a Turing-computable process.

As Searle correctly and emphatically pointed out, it does not follow from the mere fact that some state of affairs can be correctly described or simulated in digital computational terms, that it strictly encodes, inherently implements, or *really incorporates* digital computation.[73] Indeed, virtually anything in the actual physical world can be correctly described or simulated in Turing-computable terms. But it does not follow from the mere fact that a heap of empty cans of Dale's Pale Ale, or the number of steps I must take in order to reach the door of this room, could indeed be correctly described or simulated in Turing-computable terms, that either this heap or my walking across the room really incorporates a Turing-computable process. Similarly, but even more radically, far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic systems such as the roiling movements of boiling water, the paths taken by falling leaves, and weather systems, not to mention the Belousov-Zhabotinsky chemical oscillation reaction, under certain catalytic conditions with light excitation,[74] and living organisms, can indeed be correctly described or simulated in digital computational terms, but they do not *really incorporate* Turing-computable processes, precisely because they are uncomputable processes.

What is the essential difference, then, between a Turing-computable process and an uncomputable process? Again non-technically put, an essential feature of a Turing-computable process is the fact that, for each given stage in the process, there is a sufficient logical or mathematical reason why the process will

> either (i) "halt" or stop there,
> or else (ii) not "halt" or stop there.

This is because every effectively decidable procedure is inherently what is known as a *terminating process*. By sharp contrast, then, an uncomputable process is such that, for each given stage in the process, there is a sufficient reason why the process will fail to comply with any terminating Turing-computable algorithm. Uncomputable processes are therefore inherently *non-terminating processes*.[75] That an uncomputable process is "non-terminating" means that it is computationally interminable, in the sense that the process will necessarily always go on and on, unless there is a sufficient non-computational reason for it to halt or stop at some point either prior to or else beyond the halting or stopping step of any terminating Turing-computable algorithm—in which case, its halting or stopping is due to some sufficient factor within that process, whose operations are inherently beyond Turing-computation. In either case, a non-terminating process in the natural world *essentially requires some actual far-from-equilibrium, complex, self-organizing thermodynamic development in asymmetric space and time, especially time*, that cannot be pre-determined by any Turing-computable algorithm. As they say, time is of the essence.[76]

This line of thinking, in turn, allows me to formulate three closely related notions: first, that of what I call *a non-naturally-mechanistic system*, second, that of a *naturally purposive or naturally teleological system*, and third, that of what I call an *intentionally active system*.

First, X is a non-naturally-mechanistic system if and only if

> neither the structural relationships between the proper parts of X nor the causal powers of X are necessarily determined by the general deterministic or indeterministic causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, together with some Turing-computable algorithm or recursive function, but instead are necessarily determined by some sufficient factor, internal to the system, that is inherently non-Turing-computational in nature, and essentially bound up with some actual non-equilibrium, asymmetric, complex thermodynamic development in space and time.

A paradigmatic example of non-naturally-mechanistic systems is the Belousov-Zhabotinsky chemical oscillation reaction.[77]

Second, X is a naturally purposive or naturally teleological system if and only if

> (i) *X* is a non-naturally-mechanistic system, and
> (ii) *X* is also *self-organizing* in that
> (iia) the proper parts of *X* are efficient causes of each other, and
> (iib) *X* as a whole is the formal and final cause of its proper parts.

A paradigmatic example of naturally purposive or naturally teleological systems is a living organism.

Third, *X* is an *intentionally active system* if and only if

> (i) *X* is a naturally purposive or naturally teleological system, and
> (ii) *X* contains within itself a consciously-presented representational content *Y* with fine-grained normative attunement, such that for every possible Turing-computable algorithm/recursive function *Z* that describes the formal/structural relationships between the proper parts of *X* or the causal powers of *X*, *Y* can tell *X* either (iia) to "halt"/stop *prior to* the terminating step of *Z*,
> or else (iib) to go on *beyond* the terminating step of *Z*.

A paradigmatic example of intentionally active systems is a minded human animal, whether non-rational or rational.

In other words, then, in addition to their being non-naturally-mechanistic systems, naturally purposive or naturally teleological systems also exhibit the uncomputable fact of *self-organization*, or reciprocal formal and efficient causality between proper parts and whole. And in addition to teir being non-naturally-mechanistic, naturally purposive or naturally teleological systems, intentionally active systems also exhibit the further uncomputable facts of

> (i) conscious intentionality, or inherent self-guidedness by consciously-presented representational contents (either conceptual or essentially non-conceptual[78]),
> (ii) finegrained normative attunement, or inherent self-guidedness by many-degreed evaluative standards of success or failure, and
> (iii) causal spontaneity, or self-guided efficacious sourcehood, strictly underdetermined by settled facts about the past or prior event-causes.

As we shall see later in this chapter, my notion of a naturally purposive or naturally teleological system bears an important similarity to what Kant calls a "natural purpose" or *Naturzweck*, and my notion of an intentionally active system also bears a close resemblance to what Kant calls an "animal":

> animals [like humans] also act in accordance with representations[79] (and are not, as Descartes would have it, machines), and in spite of their specific difference, they are still of the same genus as human beings (as living beings). (*CPJ* 5: 464, underlining added)

Now the notions of a naturally purposive or naturally teleological system and of an intentionally active system are abstract specifications of certain kinds of thermodynamic processes. In the actual natural world, at least some[80] of the naturally purposive or teleological systems are *living systems* or organisms. Correspondingly, all of the intentionally active systems are at once *conscious systems*, or minded animals, and also

> either (i) *free volition systems*, that is, minded animal agents as such,
> or (ii) *free will systems*, that is, rational minded animal agents.

It is important to note, however, that *not all* non-naturally-mechanistic systems are naturally purposive or naturally teleological (for example, the Belousov-Zhabotinsky reaction, without a catalyst or light-excitation, and various kinds of quantum-mechanical phenomena[81]); that *not all* naturally purposive or teleological systems are also living systems or organisms (for example, the roiling movements of boiling water, the paths of falling leaves, and weather

systems); and also that *not all* living systems or organisms are also intentionally active (for example, unicellular organisms, fungi, and plants). But *all* intentionally active systems or minded animals are naturally purposive and alive, in addition to being conscious. Moreover, *not all* free volition systems or minded animal agents also have free will (for example, bats, cats, and rats). Nevertheless, *all* free will systems or rational minded animal agents are alive, conscious, and have free volition, in addition to possessing capacities for self-consciousness, conceptualization, judgment, judgment, logical reasoning, and practical reasoning (for example, us). And *all* such free will systems are intentionally active systems, naturally purposive or teleological systems, and non-naturally-mechanistic systems.

For the philosophy of free agency, obviously the leading type of intentionally active, naturally purposive or naturally teleological system is *rational human minded animal agents*—in short, *us*. We are far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, conscious, intentional, caring, rationality-guided, free-willed, practically agential thermodynamic systems within the human species, that implement all sorts of uncomputable processes, and *therefore* we are not naturally mechanized.

To make this crucial conceptual point more vivid, I want to borrow and update Leibniz's famous argument for anti-mechanism in the *Monadology*,[82] later echoed by Searle's equally famous Chinese Room argument against the Turing-test-inspired thesis of Strong Artificial Intelligence,[83] in the following way. Just to give this neo-Leibnizian, neo-Searlian argument a convenient label, I will call it "The Handwaving Room Argument":

> The Handwaving Room Argument
>
> (1) Conceive of a room which is either itself an operative real-world Universal Turing Machine, or contains inside itself an operative real-world Universal Turing Machine.
> (2) Conceive of a rational human minded animal or real human person inside that room, who is following each and every distinct digital processing movement of the Turing Machine, step-by-step with a uniquely corresponding intentional movement of her body (say, an intentional handwaving movement in which her hand forms a sequence of "zero" or "one" shapes with her fingers).
> (3) Suppose that the Turing Machine halts because its digital processing sequence has mathematically terminated.
> (4) Conceive that the real person would have been able either (i) spontaneously to stop the process prior to that terminating handwaving step in the sequence, and thus not perform that handwaving movement, or else (ii) spontaneously to go on beyond that terminating handwaving step and perform a completely different handwaving movement, in either case just because she feels like doing it at that very moment.
> (5) Now generalize and conceive that for each and every terminating process of the real-world Universal Turing Machine, at the point of termination, there is a possible real human person who has been mirroring that mathematical process step-by-step with her intentional body-movements, who would have been able either (i) spontaneously to stop the process prior to the terminating step and thus not perform that body-movement, or else (ii) spontaneously to go on beyond that step and perform a completely different body-movement, in either case just because she feels like doing it at that very moment.

> (6) Therefore at least some of the body-movements of rational human minded animals or real human persons are not only non-naturally-mechanistic processes, but also naturally purposive or teleological processes and intentionally active processes.
> (7) Therefore, since we really and truly exist in the natural world, then there really and truly are some naturally purposive or teleological systems, and also some intentionally active systems, existing in the natural world.

It is directly relevant to note in this connection that if we dropped the plausible assumption that the causal powers of any real-world Turing machine are held fixed under our general deterministic or indeterministic causal laws of nature, and if real-world Turing machines could radically vary their causal powers, then it seems that then there would be no fundamental physical, mathematical, or metaphysical difference between Turing-computable and Turing-uncomputable functions; and correspondingly it seems that then there would be no fundamental physical, mathematical, or metaphysical difference between machines and non-machines, including all naturally purposive or teleological systems and all actual-world living organisms.[84] But this claim, I think, is just equivalent to a philosophically interesting but not at all exciting thesis to the effect that if some real-world Turing machines, *contrary to actual fact, and perhaps even necessarily contrary to actual fact*, were self-organizing thermodynamic systems, then there would be no fundamental physical, mathematical, or metaphysical difference between machines and non-machines, and ultimately no deep difference between real-world Turing machines on the one hand and naturally purposive or teleological systems or actual-world living organisms on the other. Here is an analogy: Suppose it is true that if apples were changed into oranges by sending crates of apples into Malament-Hogarth spacetime,[85] then you could make orange juice out of apples. That is philosophically interesting, but not at all metaphysically exciting, since we have no reason whatsoever to think that it is actually true that apples can be changed into oranges by sending them into Malament-Hogarth spacetime. Indeed, for all we know, it is logically or strongly metaphysically *impossible* that apples can be changed into oranges by sending them into Malament-Hogarth spacetime; and since any statement whatsoever follows from a necessary falsehood, counterfactual statements with impossible antecedents are all vacuously true.

It is also directly relevant to note in this connection that if we dropped the plausible assumption that the "digits" over which the Turing machine computes are all denumerable sets of spatiotemporally discrete physical objects, and if some effectively deciding or recursive machines could compute over non-denumerable sets (for example, real, complex, or transfinite number quantities) of non-discrete (that is, either continuous or vaguely-bounded) physical items, then it seems that the Church-Turing thesis would be false, in the sense that there would then be some effectively decidable procedures or recursive functions in real physical nature which are not *classically* Turing-computable.[86] But this claim, I think, is just equivalent to another philosophically interesting but not at all exciting thesis, this time to the effect that that if some set of items over which some effectively deciding or recursive machine computes, *contrary to actual fact, and perhaps even necessarily contrary to actual fact*, were just like non-denumerable sets of non-discrete neural assemblies in the human brain, then our brains would be real physical computing machines that are not digital. Here is another analogy: Suppose it is true that if apples were just like non-denumerable sets of non-discrete neural assemblies in the human brain, then you could make orange juice out of apples. Again, that is

philosophically interesting, but not at all metaphysically exciting, since we have no reason whatsoever to think that it is actually true that apples are just like non-denumerable sets of non-discrete neural assemblies in the human brain. Indeed, for all we know, it is logically or metaphysically impossible that apples are just like non-denumerable sets of non-discrete neural assemblies in the human brain; and, again, since any statement whatsoever follows from a necessary falsehood, this guarantees that counterfactuals with impossible antecedents are vacuously true.

So Natural Mechanism says that all the causal powers of everything whatsoever in the natural world are ultimately fixed by what can be digitally computed on a universal deterministic or indeterministic real-world Turing machine, provided that the following three plausible "causal orderliness" and "decompositionality" assumptions are all satisfied:

> (i) its causal powers are necessarily determined by the general deterministic or indeterministic causal natural laws, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang,
> (ii) the causal powers of the real-world Turing machine are held fixed under our general causal laws of nature, and
> (iii) the "digits" over which the real-world Turing machine computes constitute a complete denumerable set of spatiotemporally discrete physical objects.

In direct opposition to Natural Mechanism in this precisified sense, however, the general thesis of *anti-mechanism*, as I am understanding it, says that the causal powers of biological life (and in particular, the causal powers of living organisms, especially including rational human minded animals) are neither fixed by, identical with, nor otherwise reducible to—and in particular, neither strongly nor logically supervenient on—Conservation-Law-determined, Big-Bang-caused, real-world-Turing-computable causal powers of thermodynamic systems, whether these causal powers are governed by general deterministic laws or general probabilistic/statistical laws. So if the general thesis of anti-mechanism, as I am understanding it, is true, then Natural Mechanism is false.

2.3 KANT'S ANTI-MECHANISM, KANTIAN ANTI-MECHANISM, VITALISM, AND EMERGENTISM[87]

It is well-known that in the *Critique of Pure Reason*, the *Prolegomena to Any Future Metaphysics*, and especially the *Metaphysical Foundations of Natural Science*, Kant is a self-described *Newtonian mechanist* about the manifest natural spacetime world, in which, as human animals, we must live, move, and have our being. But as early as 1763, in his pre-Critical or Leibnizian/Wolffian period, in "The Only Possible Argument in Support of a Demonstration of the Existence of God," Kant explicitly rejected the preformationist conception of biological generation and embryogenesis, according to which creatures pre-exist in their basic forms or structures, and require only the mechanical addition of bulk in order to develop. Instead, he defended the *epigenetic* view, whereby the basic forms or structures of creatures themselves are emergently generated by the spontaneous but also rule-governed operations of a goal-oriented or teleological vital source of some kind. He even went so far as to assert that:

> it would be absurd to regard the initial generation of a plant or an animal as a mechanical effect incidentally arising from the universal laws of nature. (*OPA* 2:114)

In the *Prolegomena* he asserted the identity (or at least the strong continuity) of mind and life: "life is the subjective condition of all our possible experience" (*Prol 4*: 335). In the Introduction to *Metaphysical Foundations*, he denied that there could ever be a naturally mechanistic science of psychology (*MFNS* 4:471). In the second half of the *Critique of the Power of Judgment*, he not only asserted that "the mind is for itself entirely life (the principle of life itself)" (*CPJ* 5: 278) and also that

> it would be absurd for humans ever to … hope that there might yet arise a Newton who could make comprehensible even the generation of a blade of grass according to natural laws (*CPJ* 5: 400),

but also worked out a number of fundamental concepts and methodological themes in the philosophy of biology, including the notion of a living organism, or self-organizing system, the various distinct kinds of teleology, and the special role of teleological concepts and teleological thinking in the natural sciences. And finally, in the unfinished "Transition" project in the *Opus postumum*, Kant also hypothesized the dual emergence of natural mechanisms and organismic life (including mind) alike from a single ontologically neutral but also non-static material substrate, the dynamic aether (*OP* 21: 206-233, and 241).

So Kant's commitment to Newtonian mechanism is, at the very least, somewhat conflicted. Indeed, it is fully arguable that Kant is ultimately an *anti-mechanist*. This, in turn, is the upshot of Jennifer Mensch's fascinating philosophical-historical study, *Kant's Organicism*:

> [*Kant's Organicism*] starts by tracing the history of the life sciences as Kant would have come to know them, focusing especially on those philosophers and life scientists whose works directly engaged Kant during his intellectually formative years. Once Kant's connection to the life sciences has been established, the remainder of the book moves to an examination of the exact nature of the influence of these sciences on the emerging critical system. When viewed from the perspective the life sciences in this manner, Kant's theoretical philosophy becomes reframed as a philosophical project whose development was deeply influenced by the rise of organicism.[88]

In Mensch's terminology, the thesis of *organicism*, in turn, "can be defined by its view of nature as something that cannot be reduced to a set of mechanical operations."[89] So what she calls "organicism" is essentially equivalent to what I have been calling "anti-mechanism."

Amongst other things, *Kant's Organicism* carefully describes the intellectual state-of-play in natural history in the 17th and early 18th centuries. The first players are the mechanist corpuscularian Boyle, and Locke:

> Locke was both a nominalist regarding species determination and a realist in believing that there were inner features contributing to species as well. In a similar fashion, Locke was both comfortable with a mechanical portrait of animal functioning and cognizant of the need for "inner principles" and "transformative forces" when it

> came to understanding the processes of organic life. And all this contributed to Locke's views of both nature and the proper task of classification. Reviewing Locke's early considerations of organic processes aginst the backdrop of corpuscular ontology reveals his sensitivity to the problems facing Boyle in the case of organic life. While Locke remained committed to the essential features of corpuscular science, he was nonetheless hesitant in the face of a straighforward endorsement of mechanical accounts of generation.[90]

A similar hesitation as between mechanism and anti-mechanism can be found in the work of the second major player, Leibniz, who, heavily influenced by the Dutch microscopist Leeuwenhoek, took the view that "individuals were composed of living monads arranged hierarchically under a dominant entelechy or soul."[91] In the *Monadology*, anticipating both the Turing Test and also Searle's Chinese Room argument, Leibniz famously argued, by means of a thought-experiment whereby the goal-directed conscious processes of mind cannot be reduced to the external behaviors of an enormously complicated mill, that mentality cannot be reduced to physical mechanical operations. But at the same time, Leibniz also thought of the living monads as *spiritual automata* pre-programmed by a 3-O (namely, Omniscient, Omnipotent, and Omnibenevolent) God, the supreme monad, and endorsed preformationism.

One philosophical moral of this part of the story, I think, is that the very idea of natural mechanism is typically a conceptual hybrid that combines these three sub-ideas:

> (i) *necessitation under general causal natural laws*, especially the Conservation Laws, which guarantee causal-nomological closure with respect to quantities of matter and/or energy,
> (ii) Turing-computability, and
> (iii) universal natural determinism.

But as I pointed out in section 2.2,

> (iv) there is good reason to enrich the typical concept of natural mechanism so as also to include the real possibility of *natural indeterminism* under its rubric, and
> (iv) although necessitation under general causal natural laws, especially the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, is sufficient for Turing-computability and determinism, it is *not*, strictly speaking, necessary.

For according to the Leibnizian account, non-physical automata are at least conceptually and weakly metaphysically possible. Therefore, we need to distinguish carefully between

> (1) *causal-nomological mechanisms* (for example, Coke machines), which are, necessarily, physical, and
> (2) *formal mechanisms* (for example, Turing-computable processes), which, although they are often physically realizable, are *not always physically realizabl*e (for example, in cases in which the real-world Turing machine would have to be bigger than our physical universe), and which, above all, *are not necessarily physical*: in principle, disembodied Cartesian souls could run Turing-computable sequences.

Kant is at least implicitly aware of this important distinction between causal-nomological mechanisms and formal mechanisms, because in the *Critique of Practical Reason* he explicitly rejects the reduction of all spontaneous activity, including organismic life, but also especially including free will, to the operations of Leibnizian spiritual automata, deriding the latter as "the freedom of a turnspit" (*CPrR* 5: 97).

Mensch also traces the origins of organicism to Georges Buffon's highly influential epigenesist treatise, *Natural History*, the first three volumes of which appeared in 1749:

> With Buffon natural history … became an attempt to grasp a living nature, to grasp species across time and, as a consequence, to base the classification of species upon genealogy. This marked a dramatic transformation in the history of a discipline that until then had been first and foremost a science oriented by its search for the means of discovering nature's divisions and, for that reason, not at all by the patterns of its underlying unity.[92]

Strictly speaking, Buffon's version of epigenesis is still compatible with mechanism (whether causal-nomological or formal). Correspondingly, the full theory of epigenesis would have to await the further postulation, in the 1780s, of organic or emergent vital forces, "like Caspar Wolff's *vis essentialis* and Johan Blumenbach's *Bildungstrieb*"[93]—which of course anticipate later, more famous 19th and 20th century vitalist metaphysical notions like Schopenhauer's *Wille zum Leben* and Bergson's *élan vital*. Nevertheless, the ground was prepared for Kant's organicism.

Mensch also provides a good account of Kant's pre-Critical work on cosmological and biological questions of origin, and shows how this work not only smoothly fused with, but also primed, his Critical concern with the origins, scope, and limits of cognition and knowledge. As Mensch puts it:

> [There was] an intimate connection, in Kant's view, between attempts to discover a "principle of life" within natural organisms and the search for something beyond the limits of the everyday world.[94]

In other words, Kant found a paradigm case of the burning need for his Critical distinctions between *phenomena and noumena* on the one hand, and between *the transcendental and the empirical* on the other hand, in the debate about the origins of life:

> It was the unity of purposes within organic life, the fact that organisms could be both self-sustaining and vigilant regarding the need for repair, that made natural products amazing, not the mechanical operations themselves. For Kant it was thus the principle of life, the capacity for a being's generation and self-organization that needed explaining, and recourse to neither supernatural nor purely mechanical grounds of explanation could satisfy that need.[95]

Basically, what is humanly cognizable and knowable about life (I will call these "the organicist or anti-mechanical phenomena") are the non-mechanical, spontaneous activities of the perceivable organism: but these are *not* the inherently mysterious activities of some vital

substance with an intrinsic non-relational essence, subsisting outside of manifest space and time, hiding behind the appearances (I will call this "the organicist or anti-mechanical noumenon").

Kant's organicism, as Mensch's book very effectively shows, captures Kant's Leibniz-inspired deep insight that, even when we methodologically bracket out all epistemic or metaphysical consideration of noumena, natural-mechanical principles and facts cannot explain the basic organicist or anti-mechanical *phenomena*, including

> (i) natural teleology or organismic life, including plants and animals,
> (ii) any organism with proprioceptive, enantiomorphic awareness of the difference between its right side and its left side (or top and bottom, or front and back, etc.), or an awareness of the difference between its own past, present, and future: the feeling of egocentrically-centered embodied orientation in a global space-structure with intrinsic directions ("here"), and an egocentically-centered asymmetric duration in a global time-structure ("now"), that is, *the feeling of organismic, conscious life*, whose phenomenal characters are all modes of pleasure or pain,
> (iii) human mentality, including consciousness, intentionality, imagination, conceptualizing, judging, and inferential reasoning,
> (iv) human spontaneity, agency, and source-incompatibilist free will, and
> (v) human non-instrumental normativity.

But at the same time, Kant himself could never fully advance beyond the thesis that organicist/anti-mechanist concepts have only a *regulative* (that is, methodological) use, not a *constitutive* (that is, objectively real) use.

Why not? I think that he was needlessly bedazzled by the very ideas of Newtonian mechanics and Newtonian deterministic natural mechanism, as jointly constituting a hyper-successful research program in 17th and 18th century natural science. *Over*-impressed by this (admittedly still very impressive) Newtonian program, Kant could not see that the existence of a manifest natural world that fundamentally contains significantly many causal-nomological-mechanical and formal-mechanical deterministic processes is perfectly consistent with the equally manifest organicist fact that the natural world *also* fundamentally contains significantly many *non-mechanical*, *non-deterministic* processes in it, including teleological and mental processes, as well as inherent non-instrumentally normative rules guiding these processes. Indeed, we already know from Gödel's second incompleteness theorem that formal-mechanical processes of Turing-computable *proof* presuppose non-mechanical semantic processes of non-Turing-computable *truth-determination*. So universal *formal* mechanism is provably false. Why then should we accept universal *causal-nomological* mechanism, especially when one of its necessary conditions is the supposed universality of formal mechanism?

In other words, what I am proposing is that, from a Kantian point of view, with the organicist or anti-mechanical phenomena as a starting-point, we can metaphysically postulate that the natural world is fundamentally *dual aspect*, and that it is at once mechanical-deterministic in one of its fundamental dual aspects, and also non-mechanical-non-deterministic (and also non-*in*deterministic, although Kant himself does not make this point, living and working, as he did, a century or so before "the emergence of probability"[96]) in the other of its fundamental dual aspects, including the irreducible existence of both causally-

nomologically non-mechanical processes and also formally non-mechanical processes. So, quite apart from Kant's own needless (although in charitable retrospect, perfectly understandable) deference to the Newtonian research program, we can now, in a fully Kantian spirit, put forward the radical thought that there is a fully *constitutive* use of organicist or anti-mechanical concepts, *insofar as they are required by a transcendental inference to the best explanation of all the organicist or anti-mechanical phenomena*.

And contemporary Kantians are not the only ones making such a proposal. For example, Thomas Nagel formulated essentially the same point in *Mind and Cosmos* (for which, not unsurprisingly, he received a torrent of angry criticism from scientific naturalists, not least because of his frank admission that in that book he was, in effect, proposing a version of "objective idealism"[97]), by asserting that in order to make progress on the leading problems in contemporary philosophy of mind—including how to explain the mind-body relation, mental causation, freedom, and the nature of rationality—we must metaphysically postulate a "cosmic predisposition to the formation of life, consciousness, and the value that is inseparable from them."[98]

In any case, here is the basic line of reasoning behind this radical contemporary Kantian thought. Kant's fundamental philosophical problem, the one that he struggled with throughout his long philosophical career, is this:

> How can the existence of non-mechanical, non-deterministic facts that are necessary for the purposes of morality, be made consistent and coherent with the thesis that necessarily, all the natural objects studied by physics (namely, the "objects of experience") are mechanical and deterministic?

Since all organisms, including conscious rational human organisms, or human persons, are non-mechanical and non-deterministic, then Kant's fundamental problem can be focused like a laser beam on this specific formulation of his fundamental problem:

> How can the existence of living conscious rational human animals, that is, human persons, capable of genuine incompatibilistic free will, necessary for the purposes of morality, be made consistent and coherent with with the thesis that necessarily, all the natural objects studied by physics (the "objects of experience") are mechanical and deterministic?

Now as every reader of the first *Critique* knows, for Kant, there are two basic kinds of objects:

> (i) *phenomena*, namely spatiotemporal objects directly accessible to and knowable by human sensory intuition and sense perception, that are constituted by relational properties, especially including relations to actual or possible human sensible minds, and
> (ii) *noumena*, namely non-spatiotemporal, humanly sensorily inaccessible, unperceivable, and unknowable objects, which may or may not exist, but even if they do exist, are constituted by intrinsic non-relational properties, and are at best barely consistently thinkable by means of concepts, and neither cognizable nor knowable.

But what many readers of the first *Critique* have *not* noticed is that equally important for Kant is the distinction, exclusively within the domain of phenomena, between

> (ia) *undetermine*d objects of empirical intuition, aka *appearances* (*CPR* A20/B34), and
> (ib) *fully determined* objects of empirical intuition, falling under empirical concepts, empirical judgments, and above all, falling under pure *a priori* concepts of the understanding, or Categories, aka *objects of experience* (*CPR* B161).

For Kant, as a Newtonian mechanist and also a LaPlacean determinist about physical nature insofar as it is correctly described by physics, *mechanism necessitates natural determinism*, and conversely, *natural determinism entails mechanism*. So all the actual and possible objects of experience are mechanical and deterministic. But here's the rub: all and *only* the actual and possible objects of experience are mechanical and deterministic, but *not* all the actual or possible appearances. Since the total set of pure *a priori* concepts of the understanding, or Categories, specifies a world of objects inherently governed by Newtonian mechanistic principles and laws, then, although all the fully determined objects, namely, the objects of experience, are inherently governed by Newtonian mechanistic principles and laws, and therefore are *deterministic* and not free, it does *not* follow that all or indeed any of the *un*determined objects, namely, the appearances, are either mechanical (whether causally-nomologically-mechanical or formal-mechanical) or deterministic. In other words, since for Kant the sensible intuitability of an object, independently of concepts, is the criterion of the object's real possibility, then it is either actual or at least really possible that at least *some appearances* are non-mechanical and non-deterministic, and that they are cognitively accessible by means of *essentially non-conceptual sensible intuitions*.[99]

Let us call such essentially non-conceptually sensibly intuitable appearances, insofar as they actually exist, or were they to exist, *rogue objects*, since they fall outside the Categories and the system of transcendental principles, or at least fall outside Kant's "constitutive" *causal-dynamical principles* (namely, the Analogies of Experience, and the Postulates of Empirical Thought) and therefore outside the deterministic causal laws of nature,[100] even if they do continue to fall under the "regulative" *mathematical principles* (namely, the Axioms of Intuition, and the Anticipations of Perception). The actual existence or real possibility of rogue objects would mean that the phenomenal natural world, namely, the manifest world, namely, the world of Sellars's Manifest Image, actually or really possibly includes some *appearances* that are also not *objects of experience*, namely the rogue objects, and that we can access these rogue-object phenomena only through essentially non-conceptual intuition. These non-mechanical, non-deterministic rogue-object phenomena, in turn, would include *all the organicist or anti-mechanical phenomena*, as specified above, and this would in turn directly imply that the phenomenal natural or manifest world includes some objects that are also *not* objects of mechanistic physics, mechanistic chemistry, and mechanistic biology, and therefore also that mechanistic natural science is *not*, to borrow Sellars's phrase, "the measure of all things."[101] So *scientific* or *physicalist* naturalism (whether reductive or non-reductive) would be false, and mechanistic natural science would apply to all and only the natural objects and facts to which it applies, but not to all actual or possible natural objects and facts. In short, mechanistic natural science would have *philosophical limits encountered inside nature itself*.

Contrary to scientific or physicalist naturalism, then, the thesis of *liberal* naturalism would be true. Again, the liberal naturalist thesis says

> (i) that the manifest world fundamentally contains the real existence and/or real possibility of organismic life, the feeling of life, mind, source-incompatibilist free will, persons, and non-instrumental normativity as basic organicist, anti-mechanist facts of nature, along with the basic formal-mechanical and causally-nomologically-mechanical physical facts, and
> (ii) that the basic kind of item is *thermodynamic systems, or thermodynamic processes*, both mechanical/deterministic and non-mechanical/non-deterministic, such that
> (iii) the mechanical/deterministic kind presupposes either the actual existence or the real possibility of the non-mechanical, non-deterministic kind.

Bluntly put: *source-incompatibilist free will is a fact of organismic life, and partially constitutive of physical nature*. Or in Nagel's words again, "rational intelligibility is at the root of the natural order," and there is a "cosmic predisposition to the formation of life, consciousness, and the value that is inseparable from them." This, in turn, would solve Kant's fundamental problem, not by appealing to anything *super*natural, but instead *by liberalizing our concept of physical nature*.

Anti-mechanism in its classical early 20th century guise, as "British emergentism," has its original intellectual roots in Aristotle's *De Anima* and *Physics*, and in the 17th and 18th century epigenesist-organicist tradition so well described by Mensch, when these accounts are combined with late 18th and early 19th century Romantic conceptions of nature, expressed for example in the seventh of Rousseau's *Reveries of a Solitary Walker*, Wordsworth's and Percy Shelley's poetry, and their notion of "natural piety," by Mary Shelley's stunning critique of mechanistic-reductive scientific sins against natural piety, in *Frankenstein*, and by Caspar David Friedrich's and J.M. Turner's nature paintings. All or most of these, in turn, have their proximal intellectual sources in Kant's assertions of the cognitive-semantic limits of science and scientific knowledge in the *Critique of Pure Reason*, of anti-mechanism in his moral and political philosophy, and also of a direct epistemic, metaphysical, and moral link, via immediate consciousness, between the "starry heavens above me" and the "moral law within me" at the end of the *Critique of Practical Reason*, taken together with his closely-related notions of *the beautiful in nature*, *the sublime*, *genius*, *life*, and *purposiveness-without-a-purpose* in the *Critique of the Power of Judgment*. Correspondingly, here are some of the most important texts in this "natural piety" tradition, running from Rousseau and Kant through Wordsworth, and the Shelleys to the British emergentist, Samuel Alexander:

> A deep and sweet revery seizes your senses, and you lose yourself with a delicious drunkenness in the immensity of this beatiful system with which you identify yourself. Then all particular objects fall away; you see nothing and feel nothing except in the whole… I never meditate or dream more delightfully than when I forget my self. I feel indescribable ecstasy, delirium in melting, as it were, into the system of beings, in identifying myself with the whole of nature.
>
> Brilliant flowers, enamelled meadows, fresh shades, streams, woods, verdure, come, purify my imagination … My soul, dead to all strong emotions, can be affected

now only by sensory objects, and it is only through them that pleasure and pain can reach me.[102]

[I] had to deny **scientific knowing** (*Wissen*) in order to make room for **faith** (*Glauben*). (*CPR* Bxxx, boldfacing in the original)

When nature has unwrapped, from under this hard shell, the seed for which she cares most tenderly, namely the propensity and calling to *think* freely, the latter gradually works back upon the mentality of the people (which thereby gradually becomes capable of *freedom* in acting) and eventually even upon the principles of *government*, which finds it profitable to itself to treat the human being, *who is now more than a machine*, in keeping with his dignity. (*WE* 8: 41-42, underlining added)

All necessity of events in time in accordance with the laws of natural law of causality can be called the *mechanism* of nature…. Here one looks only to the necessity of the connection of events in a time series as it develops in accordance with natural law, whether the subject in which this development takes place is called *automaton materiale*, when the machinery is driven by matter, or with Leibniz *spirituale*, when it is driven by representations; and if the freedom of our will were none other than the latter…, then it would at bottom be nothing other than the freedom of a turnspit, which, when once it is wound up, also accomplishes its movements of itself. (*CPrR* 5: 97, underlining added)

[T]wo things fill the mind with ever new and increasing admiration and reverence, the more often and more steadily one reflects on them: *the starry heavens above me and the moral law within me*. I do not need to search for them and merely conjecture them as though they were veiled in obscurity or on the transcsndent region beyond my horizon; I see them before me and connect them immediately with the consciousness of my existence. (*CPrR* 5: 161-162, underlining added)

An organized being is … not a mere machine, for that has only a motive power, while the organized being possesses in itself a formative power, and indeed one that it communicates to matter, which does not have it (it organizes the latter): thus it has self-propagating formative power, which cannot be explained through the capacity for movement alone (that is, mechanism). (*CPJ* 5: 374)

It is quite certain that we can never adequately come to know the organized beings and their internal possibility in accordance with merely mechanical principles of nature, let alone explain them; and this is so certain that we can boldly say that it would be absurd for humans to make an attempt or to hope that there could ever arise a Newton who could make comprehensible even the generation of a blade of grass according to natural laws that no intention has ordered; rather we must absolutely deny this insight to human beings. (*CPJ* 5: 400, underlining added)

My heart leaps up when I behold
A rainbow in the sky:
So was it when my life began;
So is it now I am a man;
So be it when I shall grow old,

Or let me die!
The Child is father of the Man;
And I could wish my days to be
Bound each to each by natural piety.[103]

Earth, ocean, air, belov'd brotherhood!
If our great Mother has imbued my soul
With aught of natural piety to feel
Your love, and recompense the boon with mine.[104]

One of the phaenomena which had peculiarly attracted my attention was the structure of the human frame, and, indeed, any animal endued with life. Whence, I often asked myself, did the principle of life proceed? ….To examine the causes of life we must first have recourse to death. I became acquainted with the science of anatomy: but this was not sufficient; I must also observe the natural decay and corruption of the human body…. Now I was led to examine the cause and progress of this decay, and forced to spend days and nights in vaults and charnel houses….I paused, examining and analysing all the minutiae of causation, as exemplified in the change from life to death, and death to life, until from the midst of this darkness, a sudden light broke in upon me…. After days and nights of incredible labour and fatigue, I succeeded in discovering the cause of genertion and life; nay, more, I became capable of bestowing animation upon lifeless matter…. I see by your eagerness, and the wonder amd hope which your eyes express, my friend, that you expect to be informed of the secret with which I am acquainted; that cannot be; listen patiently until the end of my story, and you will easily perceive why I am so reserved upon that subject. I will not lead you on, unguarded and ardent as I then was, to your destruction and infallible misery. Learn from me, if not by my precepts, at least by my example, how dangerous is the acquirement of knowledge, and how much happier that man is who believes his native town to be the world, than he who aspires to become greater than his nature will allow.[105]

I do not mean by natural piety exactly what Wordsworth meant by it–the reverent joy in nature, by which he wished that his days might be bound to each other–though there is enough connection with his interpretation to justify me in using his phrase. The natural piety I am going to speak of is that of the scientific investigator, by which he accepts with loyalty the mysteries which he cannot explain in nature and has no right to try to explain. I may describe it as the habit of knowing when to stop in asking questions of nature.

[T]hat organization which is alive is not merely physico-chemical, though completely resoluble into such terms, but has the new quality of life. No appeal is needed, so far as I can see, to a vital force or even an *élan vital*. It is enough to note the emergence of the quality, and try to describe what is involved in its conditions…. The living body is also physical and chemical. It surrenders no claim to be considered a part of the physical world. But the new quality of life is neither chemical nor mechanical, but something new.

> We may and must observe with care our of what previous conditions these new creations arise. We cannot tell why they should assume these qualities. We can but accept them as we find them, and this acceptance is natural piety.[106]

Now for my purposes in this chapter, Alexander's careful distinction between "vitalism," on the one hand, and "emergentism" in his sense, on the other, is crucially important. Here is a reformulation of this distinction that closely parallels the internal structure of classical Cartesian dualism in the philosophy of mind:[107]

> (i) *substance vitalism*, which says that life is an essentially different kind of dynamic stuff from naturally mechanistic matter and/or energy (for example, ectoplasm, Schopenhauer's *Wille zum Leben*, Bergson's *élan vital*, etc.), and
> (ii) *property vitalism*, or *functional vitalism*, namely, *emergentism*, which says that life is necessarily determined by essentially different kinds of dynamic functional properties from those that characterize natural mechanisms, even if life is not an essentially distinct kind of dynamic stuff from naturally mechanistic matter and/or energy.

Most of the early 20th century British emergentists were property vitalists or functional vitalists, but not substance vitalists. An essential metaphysical feature of this property vitalist, functional vitalist, or emergentist view, however, as Brian McLaughlin,[108] David Chalmers,[109] Jaegwon Kim,[110] and many others have noted, is the thesis that the irreducible functional properties and/or facts of life are *naturally or nomologically strongly supervenient* on fundamental physical properties and/or facts, with the unhappy metaphysical result that these "higher-level" biological properties are causally inert or epiphenomenal, causally excluded by the supervenience base of causally efficacious fundamental physical properties and/or facts.

It is, in particular, this singularly unhappy metaphysical feature of property vitalism or emergentism that I want to reject. And this particular rejection, in turn, sets the metaphysical view I want to defend, which I call *dynamicism*, sharply apart from natural mechanism, substance vitalism, and property vitalism or emergentism *alike*—although I do also want to preserve several basic epistemic, aesthetic, and moral features of the Romantic/British Emergentist/"natural piety" tradition in my overall views on the philosophy of nature and natural science.[111]

In any case, the basic metaphysical claim I am making in this chapter is that the new quality—better, "specific character"—of organismic life *does* indeed emerge in far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic processes *over time*, or *diachronically*, hence it is rightly called *dynamic emergence*.[112] Nevertheless, it *does not* strongly supervene on fundamental physical properties and/or facts, *at all*, but more specifically it does not strongly supervene *at-a-time*, or *synchronically*. Hence dynamic emergence is logically and metaphysically independent of any form of supervenient emergence. Leaving aside the jargon of contemporary Analytic metaphysics, my basic point is just this: organismic life arises in nature as a temporally novel, immanent structural feature of natural processes possessing a certain suitable level of thermodynamic complexity. As immanent-structural, such features do not metaphysically *pop out* of these natural processes, *at all*, but more specifically they do not metaphysically pop out *at-a-time*, or synchronically, since the thermodynamics of the processes themselves, in asymmetric time, is inherently and internally

guided and self-determined by these very same structural features. On the contrary, one less complex dynamic structure (that is, far-from-equilibrium, spatiotemporally asymmetric, and complex, but still non-self-organizing and non-living thermodynamics, for example, the Belousov-Zhabotinsky chemical reaction, without a catalyst or light-excitation[113]) *ontologically opens up and unfolds into* another, essentially richer dynamic structure (that is, organismic life), just as the less complex system of the rational-and-natural-numbers *constructively opens up and unfolds into* the essentially richer system of the real numbers by means of, for example, the power-set operation or the Dedekind-cut operation.

"Opening up and unfolding into" is of course a metaphor. It is intended to convey the basic idea that the emergence-in-and-over-time of organismic life is *essentially inside* the asymmetric spatiotemporal processes constituting its thermodynamics, just as the real numbers are *essentially between* the rational and natural numbers. Dynamic emergence, generally, is the spatiotemporally asymmetric self-revelation, unfolding, and actualization of a previously merely potential richer thermodynamic structure, with correspondingly new causal powers. Hence organismic life is neither *reducible to* fundamental physical properties and/or facts (= natural mechanism), nor is it some sort of substance ontologically *over and above* fundamental physical properties and/or facts (= substance vitalism), nor does it *pop-out emerge* from fundamental properties and/or facts (= property or functional vitalism). Organismic life reveals itself, unfolds, and efficaciously actualizes a previously merely potential richer thermodynamic structure consisting—as we shall see in the next section—of

> (i) *special teleological dynamics* (reproduction, growth, motility, death, and evolution or natural selection),
> (ii) *essential indexicality* (inherent context-dependency, centered on a frame-of-reference) and
> (iii) *causal spontaneity* (efficacious metabolism, involving DNA, and epigenesis).

It should also be added here, by way of making it obvious where I am heading with all of this, that the very same basic points about dynamic emergence are true of consciousness, caring, intentionality, essentially embodied rationality, deeply free will, practical agency, and real personhood themselves, all of which are metaphysically continuous with organismic life. None of them are naturally-mechanizable; none of them are in any way dualistic properties and/or facts; and none of them are in any way strongly supervenient on, or "pop-out emergent" from, fundamental physical properties and/or facts. Instead, they are all thoroughly immanent-structural, thermodynamic properties and/or facts that dynamically emerge, when simpler thermodynamic immanent structures successively ontologically "open up and unfold into" essentially richer thermodynamic immanent structures.

Perhaps surprisingly, at least initially, I will work my way up to the full presentation of the non-physicalist, non-dualist, non-supervenient, anti-mechanist metaphysics of dynamicism via the *cognitive semantics* of the representation of life. My full rationale for starting with a cognitive-semantic argument, as opposed to jumping right into the thick of things with a directly metaphysical argument, will become evident as we go along. But the wordbite version of the rationale is that the standard contemporary ways of doing metaphysics, by means of either conceptual analysis or modal logic, are deeply insensitive to the synthetic a priori character of *real metaphysics*, which on the contrary is not cognitive-semantically grounded on concepts or propositions, but instead cognitive-semantically grounded on essentially *non-*

conceptual representations,[114] just as it is ontologically grounded on the notions of *veridical appearances* and *manifest reality* (see section 1.0 above).

In sharp contrast to vitalism, whether substance vitalism or property/functional vitalism, Michael Thompson has argued for the following two-part thesis:

> (i) that our everyday, pre-theoretical representation of life (aka "folk biology") requires a distinctive Fregean logical form of what he calls "natural-historical judgments," and
> (ii) that this distinctive logical form entails the existence of a non-empirical concept of life with irreducible semantic content and structure, that necessarily shapes our ordinary perceptual and practical activities.

This two-part thesis, which I will call *representational anti-mechanism*, has significant anticipations and parallels in Kant's accounts of "the feeling of life," of the identity (or of at least the metaphysical continuity) of mind and life, and of teleological judgment in the *Critique of the Power of Judgment*; in the later Wittgenstein's notions of "forms of life" and "seeing-as" in *Philosophical Investigations*; and in Hans Jonas's existential philosophy of biology in *The Phenomenon of Life*. More precisely, however—and now generalizing over the several similar accounts provided by Kant, the later Wittgenstein, Jonas, and most recently Thompson—representational anti-mechanism, as I will understand it, says

> (i*) our everyday, pre-theoretical representations of life in sense perception and other essentially non-conceptual representations, in conceptual thought, and in biological or natural-historical judgments and statements, are neither necessarily determined by, nor identical with, nor otherwise reducible to naturally mechanistic theories of biology and life, and
> (ii*) these representations of life entail the existence of some a priori representations with non-physicalist, irreducible semantic content and structure, that necessarily shape our basic cognitive and practical encounters with the natural world.

Now representational anti-mechanism is fully consistent with the denials of substance vitalism and property/functional vitalism alike; moreover, if representational anti-mechanism is true, then not only, first, there is no explanatory reduction of the phenomenon of life to naturally mechanistic processes or facts, including those falling under reductive physicalist Darwinian theories, but also, second, the phenomenon of life is not strongly supervenient on any naturally mechanistic processes or facts whatsoever, including those falling under *non-reductive* physicalist Darwinian theories.

The first point closely parallels Nagel's famous "explanatory gap" argument for the irreducibility of mentalistic concepts to physicalistic concepts in "What is It Like to Be a Bat?"[115] Indeed, Nagel himself explored this close parallel in *Mind and Cosmos*. That first point also closely parallels Chalmers's well-known formulations of the *inverted qualia*, *zombie*, and *panprotopsychist* arguments for both the explanatory non-reduction and also the ontological non-reduction of consciousness to the fundamental physical world in *The Conscious Mind*—although, to be sure, this is ironic, given Chalmers's given official commitment to reductive physicalism about the phenomenon of life. More on that illuminating irony later, in section 2.6.

But the even deeper and even more important point for my purposes is the second one: *the correct cognitive semantics of the representation of life rules out any sort of physicalism, but also without entailing any sort of dualism*. Again: there is a defensible "third way," via the correct cognitive semantics of the representation of life, between the Scylla of physicalism and the Charybdis of dualism. So that is my next topic.

2.4 On the Representation of Life

As I said at the end of the last section, I think that Thompson is correct that there is a defensible argument for his two-part representational anti-mechanistic thesis, which says

> (i) that our everyday, pre-theoretical representation of life requires a distinctive logical form of biological or natural-historical judgments and statements, and
> (ii) that this distinctive logical form entails the existence of a non-empirical concept of life with non-physicalist, irreducible semantic content and structure, which necessarily shapes our ordinary perceptual and practical activities.

But I also want to hold *extended and generalized* versions of Thompson's theses, which say

> (i*) that our everyday, pre-theoretical representations of life in sense perception and other essentially non-conceptual representations, in conceptual thought, and in biological or natural-historical judgments and statements, are neither necessarily determined by, nor identical with, nor otherwise reducible to naturally mechanistic theories of biology and life, and
> (ii*) that these representations of life entail the existence of some a priori representations with non-physicalist, irreducible semantic content and structure, that necessarily shape our basic cognitive and practical encounters with the natural world.

As I also said, theses (i*) and (ii*) jointly comprise "representational anti-mechanism." Moreover, I think that representational anti-mechanism is well-supported by later Wittgenstein's remarks on "forms of life," and on seeing the difference between living things and dead things, in *Philosophical Investigations*; by recent empirical work in cognitive psychology by Deborah Kelemen on the phenomenon of "promiscuous teleology"[116]; by recent philosophical work by Tamar Gendler on the distinction between "alief" and "belief"[117]; and by Kant's accounts of "the feeling of life," of the identity of mind and life, and of teleological judgment in the *Critique of the Power of Judgment*.

Here is what later Wittgenstein says:

> Look at a stone and imagine it having sensations. —One says to oneself: How could one get so much as the idea of ascribing a sensation to a thing? One might as well ascribe it to a number! —And now look at the wriggling fly and at once these difficulties vanish and pain seems to get a foothold there, where before everything was, so to speak, too smooth for it. And so, too, a corpse seems to us quite inaccessible

> to pain. —Our attitude to the living is not the same as to the dead. All our reactions are different. —If anyone says: "That cannot simply consist in the fact that the living behave in such-and-such a way and the dead do not," then I want to intimate to him that this is a case of the transition "from quantity to quality."[118]
>
> If one sees the behaviour of a living thing, one sees its soul.[119]
>
> "To me it is an animal pierced by an arrow." That is what I treat it as; this is my *attitude* to the figure. This is one meaning in calling it a case of 'seeing'.[120]
>
> I might say: a picture does not always *live* for me while I am seeing it. "Her picture smiles down on me from the wall." It need not always do so, whenever my glance lights on it.[121]
>
> What has to be accepted, the given, is—so one could say—forms of life.[122]

Here is what Kelemen says:

> In summary, British and American children have a promiscuous tendency to teleologically explain the properties of both living and non-living things in terms of a purpose. One proposal is that this bias occurs because, during development, across cultures, children primarily develop an artifact model when reasoning about the natural world…. There are several implications if this turns out to hold truth: from a theoretical standpoint, it suggests that while teleological thought may play a crucial role in children's early reasoning about living things, its presence is not necessarily indicative of a truly "biological" [that is, physically mechanistic] mode of construal…. From an educational standpoint, it helps to explain why people consistently misinterpret natural selection as a quasi-intentional, designing force rather than as a blind physical mechanism.[123]

Here is what Gendler says:

> [Consider the following example, borrowed from an essay by Kendall Walton:] Charles is watching a horror movie about a terrible green slime. He cringes in his seat as the slime oozes slowly but relentlessly over the earth destroying everything in its path. Soon a greasy head emerges from the undulating mass, and two beady eyes roll around, finally fixing on the camera. The slime picking up speed, oozes on a new course straight towards the viewers. Charles emits a shriek and clutches desperately at his chair.
>
> How should we describe Charles's cognitive state? Surely he does not *believe* that he is in physical peril; as Kendall Walton writes, "Charles knows perfectly well that the slime is not real and that he is in no danger"…. But alongside that belief there is something else going on. Although Charles *believes* that he is sitting safely in a chair in a theater in front of a movie screen, he also *alieves* something very different. The alief has roughly the following content: "Dangerous two-eyed [living] creature heading towards me! H-e-l-p…! Activate fight or flight adrenaline now!"

> I argue for the importance of recognizing the existence of alief.… As a class, aliefs are states that we share with non-human *a*nimals; they are developmentally and conceptually *a*ntecedent to other cognitive attitudes that the creature may go on to develop. And they are typically *a*ffect-laden and *a*ction-generating.
>
> I offer the following tentative characterization of a paradigmatic alief:
>
> A paradigmatic alief is a mental state with associatively linked content that is representational, affective, and behavioral, and that is activated—consciously or nonconsciously—by features of the subject's internal or ambient environment. Aliefs may be either occurrent or dispositional.[124]

But most importantly of all, however, here is what Kant says:

> To grasp a regular, purposive structure with one's faculty of cognition (whether the manner of representation be distinct or confused) is something entirely different from being conscious of this representation with the sensation of satisfaction. Here the representation is related entirely to the subject, indeed to its feeling of life (*Lebensgefühl*), under the name of pleasure or displeasure, which grounds an entirely special faculty for discriminating and judging that contributes nothing to cognition, but only holds up the given representation in the subject to the entire faculty of representation, of which the mind becomes conscious in the feeling of its state. (*CPJ* 5: 204)

> It cannot be denied that all representations in us, whether they are objectively merely sensible or else entirely intellectual, can nevertheless subjectively be associated with gratification or pain, however unnoticeable either might be (because they all affect the feeling of life, and none of them, insofar as it is a modification of the subject, can be indifferent). (*CPJ* 5: 277)

> Life without the feeling of the corporeal organ is merely consciousness of one's existence, but not a feeling of well- or ill-being, i.e., the promotion or inhibition of the powers of life; because the mind for itself is entirely life (the principle of life itself), and hindrances and promotions must be sought outside it, though in the human being himself, hence in combination with his body. (*CPJ* 5: 278)

> For a body to be judged as a natural purpose in itself and in accordance with its internal possibility, it is required that its parts reciprocally produce each other, as far as both their form and their combination is concerned, and thus produce a whole out of their own causality, the concept of which, conversely is in turn the cause (in a being that would possess the causality according to concepts appropriate for such a product) of it in accordance with a principle; consequently the connection of **efficient causes** could at the same time be judged as an **effect though final causes**. In such a product of nature each part is conceived as if it exists only **through** all the others, thus as if existing **for the sake of the others** and **on account** of the whole, i.e., as an instrument (organ), which is, however, not sufficient (for it could also be an instrument of art, and thus represented as possible at all only as a purpose); rather it must be thought of as an organ that **produces** the other parts (consequently each produces the others reciprocally), which cannot be the case in any instrument of art, but only of nature, which provides all the matter for instruments (even those of art): only then and on that

> account can such a product, as an **organized** and **self-organizing** being, be called a **natural purpose** (*Naturzweck*) (*CPJ* 5: 373-374, boldfacing in the original)
>
> [A] mere machine … has only a **motive** power, while the organized being possesses in itself a **formative** power. (*CPJ* 5: 374, boldfacing in the original)
>
> Strictly speaking, the organization of nature is … not analogous with any causality that we know. (*CPJ* 5: 375)
>
> It might always be possible that in, for example, an animal body, many parts could be conceived as consequences of merely mechanical laws…. Yet the cause that provides the appropriate material, modifies it, forms it, and deposits it in the appropriate place must always be judged teleologically, so that everything in it must be considered as organized, and everything is also, in relation to the thing itself, an organ also. (*CPJ* 5: 377)

Now here are the five basic take-away points from these texts:

> (1) The representation of life is the representation of natural things *as living organisms*—that is, as far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic systems that inherently engage in naturally purposive or naturally teleological activities, plus some *further* special characteristic features of organisms to be described shortly.
> (2) The capacity to represent things as alive appears to be innate, in that it manifests itself in children and also more mature human cognizers under "poverty of the stimulus" conditions.
> (3) The representation of life can be *over*extended to things other than actual living organisms; but in any and every case it changes our practical attitudes towards the things that are perceived as alive or taken to be alive.
> (4) The representation of life is generated by a cognitive capacity that is "informationally encapsulated" or cognitive-semantically insensitive to contrary beliefs, and at the same time its representational outputs are presupposed by both ordinary and scientific beliefs, judgments, and thoughts about life.
> (5) As a consequence of points (1) to (4), the representation of life is non-empirical or a priori in the strong metaphysical sense that its content and structure are both necessarily and constitutively underdetermined by any and all sensory-experiential facts and/or contingent natural things or facts.[125]

And here are two further comments on Kant's theory in particular, before moving on.

First, for Kant, the representation of biological life not only has semantic content but also its own *phenomenal specific character*, which he calls "the feeling of life." This is the same as the pre-reflectively conscious *pleasure or pain* we experience in the actual operations of our cognitive faculties, insofar as they track naturally purposive or naturally teleological, organismic structure in objects and in ourselves. Kant's idea is that the semantic content of the representation of biological life and the phenomenal character of the feeling of life are necessarily mutually bound up with one another, which, if it is true, directly implies what is known in contemporary philosophy of mind as the "Phenomenology of Intentionality" and

"Intentionality of Phenomenology" theses, or "Anti-Separatism."[126] According to this Kantian picture, then, consciousness and intentionality are mutually inseparable via the neurobiological life of embodied animal minds.

Second, Kant explicitly *identifies* biological life with mind. This, I think, is best understood neither as literal identity, namely, some version of *panpsychism* with respect to biological life, according to which, necessarily everything alive is minded and conversely, nor as "downwards identity," that is, the *reduction* of mind to life. On the contrary, it is best understood as what Peter Godfrey-Smith calls "the strong continuity view":

> Life and mind have a common abstract pattern or set of basic organizational properties. The …properties characteristic of mind are an enriched version of the … properties that are fundamental to life in general. Mind is literally life-*like*.[127]

This is also what Evan Thompson calls the "mind-in-life" thesis:

> Where there is life there is mind, and mind in its most complex forms belongs to life. Life and mind share a core set of formal or organizational properties, and the formal and organizational properties distinctive of mind are an enriched version of those fundamental to life. More precisely, the *self-organizing* features of mind are an enriched version of the self-organizing features of life.[128]

In other words, mind is explanatorily and ontologically *continuous* with life, in the sense that whatever is metaphysically required for mind is also present in biological life, but not necessarily as organized in the right way and with the appropriate kind of thermodynamic immanent structure.[129] Therefore, not necessarily every living thing is conscious, but necessarily every mind is also biologically alive.

If this neo-Aristotelian and contemporary Kantian mind-in-life thesis is correct, then the way is open for thinking about conscious, intentional, caring, desiring animal minds as nothing more and nothing less than special forms of life, that grow naturally in organisms like us, and correspondingly the way is also open for thinking about phenomenology, the science of consciousness and intentionality, as nothing more and nothing less than a special branch of philosophical macrobiology—in effect, then, all phenomenology is *biophenomenology*.

What, more precisely, is the nature of the cognitive-semantic content of the non-empirical representation of life, that is, the non-empirical representation of living organisms? Using the Transcendental Aesthetic and the *Critique of the Power of Judgment* as philosophical sources, together with non-equilibrium/complex systems dynamics and contemporary biology (which I will discuss in the next section), as I mentioned earlier, I want to say that it includes three basic elements, over and above self-organization:

> (i) *special teleological dynamics in organisms*: their reproduction, growth, motility, death, and evolution or natural selection,
> (ii) *essential indexicality in organisms*: their inherent context-dependency, together with egocentric centering in a frame of reference (but not necessarily *occurrently conscious* centering—see, for example, Einstein's observer-relative frames of reference for tracking motion), together with orientable space and irreversible time (aka "time's arrow"), and

> (iii) *causal spontaneity in organisms*: their efficacious metabolism (as a matter of nomological or physical necessity, involving DNA) by means of "epigenesis."

Before pressing on, I will add a relevant follow-up comment about the third basic element. The thesis of epigenesis in biology says that biological material is initially unformed and that form gradually emerges through the non-predetermined or relatively spontaneous operations of an innate endogenous organizational or processing device in interaction with its environment.[130] Kant explicitly defends the theory that biological life is epigenetic, and also extends this theory analogically to his theory of cognitive innateness (see *CPR* B167 and *CPJ* 5: 424).[131]

2.5 Strong Kantian Non-Conceptualism and the Dynamicist Model of Life[132]

What is the cognitive-semantic nature of the non-empirical representation of life? I think that Thompson is mistaken in holding that the content of the non-empirical representation of life is *conceptual*. On the contrary, I hold that its content is *essentially non-conceptual* in the dual sense that it is both what I call *concept-independent* (that is, it is not sufficiently determined by conceptual capacities or concepts) and also *concept-autonomous* (that is, it does not require conceptual capacities or concepts as a necessary condition). I also hold that the structure of the non-empirical representation of life directly corresponds to what Kant would have called a "form of intuition" (*CPR* A19-49/B33-73).[133] As a consequence, I also think that Kant's theory of teleological judgments, when taken together with a contemporary Kantian theory of mental content that I have dubbed *Strong Kantian Non-Conceptualism*, [134] provide a significantly better account of the nature of the distinctive semantic content and structure of the representation of life than Thompson's Fregean account does. Elsewhere, I have spelled out and defended the Kantian Non-Conceptualist thesis in full detail.[135] So for the present purposes I will just briefly state it, and then continue to develop the main argument of this chapter.

The thesis of Non-Conceptualism about mental content says that not all mental contents in the intentional or representational acts or states of minded animals are necessarily or constitutively determined by their conceptual capacities, and that at least some mental contents are necessarily or constitutively determined by their non-conceptual capacities.[136] Non-Conceptualism is sometimes, but not always, combined with the further thesis that non-conceptual capacities and contents can be shared by rational human animals, non-rational human minded animals (and in particular, infants), and non-human minded animals alike. But in any case, Non-Conceptualism is directly opposed to the thesis of Conceptualism about mental content, which says that all mental contents are necessarily or constitutively determined by minded animals' conceptual capacities.[137] Conceptualism is also sometimes, but not always, combined with the further thesis that the psychological acts or states of infants and non-human minded animals lack mental content.

Now in a nutshell, Non-Conceptualism says that our cognitive access to the targets of our intentionality is neither always nor necessarily mediated by concepts, nor sufficiently determined or constituted by concepts, which is the *concept-independence* of non-conceptual content, and therefore that our cognitive access to the targets of our intentionality is sometimes

wholly unmediated by concepts, or altogether concept-free, which is the *concept-autonomy* of non-conceptual content; and Conceptualism says that our cognitive access to the targets of our intentionality is always and necessarily mediated by concepts, and indeed also sufficiently determined or constituted by concepts. Here, then, are the fundamental philosophical questions that are being asked in the debate about non-conceptual content: Can we and do we sometimes cognitively encounter other things and ourselves directly and non-discursively, hence non-intellectually or sensibly (Non-Conceptualism), or must we always cognitively encounter them only within the framework of discursive rationality, hence intellectually or discursively (Conceptualism)? Are we, as rational animals, essentially different from other kinds of animals (Conceptualism), or do we share at least some minimally basic mental capacities with all minded animals (Non-Conceptualism)? And finally: Is a thoroughly intellectualist and "discursivity first" view of the rational human mind (Conceptualism) correct, or by sharp contrast is a non-intellectualist and "sensibility first" view of the rational human mind (Non-Conceptualism) correct? I think that the "sensibility first" view is the correct one.

It is also important to note that whereas Conceptualism is of necessity a form of *content-monism*, which says there is one and only kind of intentional or representational content (sometimes, however, combined with *capacity-dualism*, which says that there are two essentially different basic kinds of cognitive capacities), by contrast Non-Conceptualism can be, and usually is, a form of *content-dualism*, which says that there are two essentially different kinds of intentional or representational content, and if so, then it is always a form of capacity-dualism too. Correspondingly, the version of Non-Conceptualism that I defend, Strong Kantian Non-Conceptualism, is both content-dualist and capacity-dualist.

As a sub-species of Non-Conceptualism, Strong Kantian Non-Conceptualism is the following three-part doctrine:

> (i) mental acts or states in minded animals have representational content whose semantic structure and psychological function are *essentially distinct* from the structure and function of conceptual content,
> (ii) the specific psychological function of essentially non-conceptual content is to guide conscious intentional desire-driven body movements for the purposes of cognition and practical agency, and
> (iii) the semantic structures of essentially non-conceptual content are equivalent to Kant's spatiotemporal forms of intuition, *including* their representational roles as "formal intuitions" (*CPR* B136), and also as non-empirical schemata or "mental models" generated by the "figurative synthesis of the imagination" (*CPR* B151-156 A137-147/B176-187).

More precisely however, according to Strong Kantian Non-Conceptualism, *X* is an *essentially non-conceptual content of perception* if and only if *X* is a mental content such that

> (i) X is not a conceptual content, as defined by a defensible, non-question-begging theory of concepts and conceptual content,
> (ii) X is included in a mental state, act, or process that directly refers to some or another causally efficacious actual individual macroscopic material being B in the local or distal natural environment of the minded animal subject of X—and it is also really possible that the minded animal subject of X = B—and thereby both

> uniquely (if not always perfectly accurately) locates B in 3D Euclidean orientable space and also uniquely (if not always perfectly accurately) tracks B's thermodynamically asymmetric and temporally irreversible causal activities in time, in order to individuate, normatively guide, and informationally mediate the subject's conscious intentional desire-driven body movements for the purposes of cognitive and practical intentional agency, and
> (iii) X is an inherently context-sensitive, egocentric or first-person-perspectival, spatiotemporally structured content that is not ineffable, but instead shareable or communicable only to the extent that another minded animal ego or first-person is in a cognitive position to be actually directly perceptually confronted by the same causally efficacious actual individual macroscopic material being B in a spacetime possessing the same basic 3D Euclidean orientable, thermodynamically asymmetric, and temporally irreversible structure.

In view of condition (ii), essentially non-conceptual content is inherently normatively governed by an ideal standard of *accurate direct reference for the purposes of location and tracking*, and can still be directly referential even when it is only more-or-less accurate. More accuracy means *better* location and tracking by the conscious subject, and less accuracy means *worse* location and tracking by the conscious subject. So in view of condition (ii), it follows that essentially non-conceptual content fully includes what Hubert Dreyfus calls "the nonconceptual world of absorbed coping," including Heideggerian "concern" and "readiness-to-hand" (that is, normatively engaged, skillful use of tools), Wittgensteinian "blind" involvement in shared practices as "forms of life," and Husserlian "lifeworld" phenomena more generally, although this must *also* be understood, as per the classical existential phenomenologists (especially Sartre and Merleau-Ponty), as normatively rich, pre-reflectively conscious content that is inherently poised for guiding the performance of basic intentional actions[138] by minded animals, and thus it is inherently *agential content*.[139]

Here is a simple argument for the existence of essentially non-conceptual perceptual content, that I will call, with backwards reference, obviously, to *The Handwaving Room Argument*, *The Handwaving Argument*. This simple argument anticipates a more complicated, explicit, and rigorously-formulated argument for the same conclusion, using the notion of directly perceivable qualitative three-dimensional material duplicates that are also mirror-reflected spatial counterparts—aka "incongruent counterparts," or "enantiomorphs," for example, (slightly idealized versions of) your right and left hands—which I spell out and defend elsewhere.[140]

> The Handwaving Argument
>
> (1) Suppose that I am standing right in front of you and saying "All bachelors are males, and all males are animals, so it is analytic that all bachelors are animals, right?" By hypothesis, you are concentrating your thoughts exclusively on what I am saying, and clearly understand it.
> (2) Suppose also that as I am saying "All bachelors are males," my arms are held out straight towards you and I am also moving my right hand, rotated at the wrist, in a clockwise circular motion seen clearly from your point of view, which is also a counterclockwise circular motion seen clearly from my point of view. By hypothesis, you are looking at this hand-movement, but not also thinking about it,

> although of course in some other context you might be looking at it and also thinking about it. But, by hypothesis, not in this context. You are seeing it but not thinking about it, just as when you are driving a car and your mind is fully focused on some train of thought having nothing to do with driving, you can see all sorts of things passing by you, and you can even skillfully drive, without thinking at all about the things that you are seeing or doing.
> (3) Suppose also that as I am saying, "…and all males are animals," I begin moving my left hand, again rotated at the wrist, in a counterclockwise circular motion seen clearly from your point of view, which is also a clockwise circular motion seen clearly from my point of view. By hypothesis again, you are also looking at this hand-movement, but not also thinking about it, although of course in some other context you might be looking at it and also thinking about it. But, again by hypothesis, not in this context.
> (4) Suppose also that as I am saying, "… so it is analytic that all bachelors are animals, right?" I am moving both hands simultaneously in front of you in the ways specified in (2) and (3).
> (5) Your conceptual capacities are being used by you to concentrate on what I am saying about bachelors, males, and animals, and to understand it clearly, which by hypothesis you do.
> (6) Insofar as you are using those conceptual capacities exclusively to concentrate on and to understand clearly what I am saying, you are not using your conceptual capacities to see clearly what I am doing with my hands.
> (7) Yet you also see clearly what I am doing with my hands. Your conscious attention in this context is divided into linguistic understanding and lucid vision, but by hypothesis in this context your conceptual capacities for linguistic understanding are neither distracted nor divided.
> (8) Therefore you are using your non-conceptual capacities to see clearly what I am doing with my hands.
> (9) The kind of mental content that individuates, guides, and mediates the use of non-conceptual capacities is essentially non-conceptual content.
> (10) Therefore essentially non-conceptual content really exists.

The Handwaving Argument is directly inspired by Kant's famous "argument from incongruent counterparts" for the truth of the thesis of the "transcendental ideality" of space and time, according to which space and time are nothing but subjective forms of human sensibility.[141] So it has Kantian historical roots. Nevertheless, although I do think that The Handwaving Argument is sound, it is not intended to be rationally decisive, because it leaves a lot of important information merely implicit—instead, it is intended to be only rationally evocative in the sense that it clearly and quickly indicates where I am heading, and primes us for what I want to argue about the representation of life.

Now there are contemporary scientific models of life,[142] fully informed by non-equilibrium thermodynamics, aka complex systems dynamics,[143] that conform much more closely to our everyday, pre-theoretical representation of life, captured by neo-Aristotelian and contemporary Kantian ideas about the representation of life, than they conform to the essentially mechanistic scientific model of life that is provided by physicalism, whether reductive or non-reductive. This is what I call *the dynamicist model of life*.

What do I mean by this, more specifically? As I noted in section 1.1 above, *dynamic systems* are unified collections of material elements in rule-governed or patterned motion. In connection with dynamic systems, *complexity* is the fact that the causally efficacious exchange of matter and/or energy between a dynamic system and its local natural environment does not remain constant, or fluctuates. Given complexity, then dynamic systems with identical, or virtually identical, initial conditions, may diverge radically over time. *Thermodynamic systems* necessarily involve energy (and degrees of energetic activity, heat), along with matter. Thermodynamic systems for which the formal structures of matter and/or energy remain the same, or essentially the same, over time, are *equilibrium, or near-equilibrium, time-reversible* systems. *Entropy* is a function of the state of a thermodynamic system that expresses the increasing amount of so-called "disorder" or "heat death" in a system, but less conceptually misleadingly, can be thought of as the increasing amount of *structural simplicity or uniformity* in a system, which rises monotonically to a maximum at equilibrium conditions. Here energy is entirely potential, without actualization or entropic motion. So equilibrium or near-equilibrium, time-reversible thermodynamic systems do not (significantly) increase entropy. By contrast, thermodynamic systems for which the structures of matter and/or energy change over time, and are temporally irreversible in that they (significantly) increase entropy but do not reach a maximum of entropy, are *far-from-equilibrium* systems. *Self-organizing* complex thermodynamic systems, in turn, are far-from-equilibrium, temporally irreversible thermodynamic systems that also have *dissipative structure* and *natural purposiveness or natural teleology*. A dissipative structure is how the increasing amount of entropy in a complex thermodynamic system is absorbed and dispersed (hence "dissipated") by the systematic re-introduction of matter and/or energy into the system, via a non-static causal balance between the inner states of the system and its surrounding natural environment. And natural purposiveness or natural teleology is how a far-from-equilibrium, temporally irreversible, complex thermodynamic system with dissipative structure self-generates forms or patterns of order that determine its own causal powers, and in turn place constraints on the later collective behaviors, effects, and outputs of the whole system, in order to maintain itself. The paradigmatic or prime example of a self-organizing complex thermodynamic system is a living organism—although, as I have mentioned above, not every self-organizing system is itself an organism.

In view of all that, here is Bruce Weber's very informative and non-technical summary description of the dynamicist model of life:

> Animate beings share a range of properties and phenomena that are not seen together in inanimate matter, although examples of matter exhibiting one or the other of these can be found. Living entities metabolize, grow, die, reproduce, respond, move, have complex organized functional structures, heritable variability, and have lineages which can evolve over generational time, producing new and emergent functional structures that provide increased adaptive fitness in changing environments. Reproduction involves not only the replication of the nucleic acids that carry the genetic information but the epigenetic building of the organism through a sequence of developmental steps. Such reproduction through development occurs within a larger life-cycle of the organism, which includes its senescence and death. Something that is alive has organized, complex structures that carry out these functions as well as sensing and responding to interior states and to the external environment and engaging

in movement within that environment. It must be remembered that evolutionary phenomena are an inextricable aspect of living systems; any attempt to define life in the absence of this diachronic perspective will be futile…. [L]iving systems may be defined as open systems maintained in steady-states, far-from-equilibrium, due to matter-energy flows in which informed (genetically) autocatalytic cycles extract energy, build complex internal structures, allowing growth even as they create greater entropy in their environments, and capable, over multigenerational time. of evolution.

The impact of Schrödinger's [*What is Life? The Physical Aspect of the Living Cell*] on a generation of physicists and chemists who were lured to biology and who founded molecular biology is well chronicled…. Knowledge about the protein and nucleic acid basis of living systems continues to be obtained at an accelerating rate, with the sequencing of the human genome as a major landmark along this path of discovery. The "self-replicating" DNA has become a major metaphor for understanding all of life. The world is divided into replicators, which are seen to be fundamental and to control development and be the fundamental level of action for natural selection, and interactors, the molecules and structures coded by the replicators…. Indeed, Dawkins relegates organisms to the status of epiphenomenal gene-vehicles, or survival machines. A reaction has set in to what is perceived as an over-emphasis on nucleic acid replication…. In particular developmental systems theorists have argued for a causal pluralism in developmental and evolutionary biology…. However, the rapid progress in gene sequencing is producing fundamental insights into the relationship of genes and morphology and has added important dimensions to our understanding of evolutionary phenomena….

What is less known is the over half-a-century of work inspired, in part, by the other pillar of Schrödinger's argument, namely how organisms gain order from disorder through the thermodynamics of open systems far from equilibrium…. Prominent among early students of such nonequilibrium thermodynamics was Ilya Prigogine…. Prigogine influenced J. D. Bernal in his 1947 lectures on the physical basis of life to start to understand both how organisms produced their internal order while affected their environment by not only their activities but through created disorder in it…. Harold Morowitz explicitly addressed the issue of energy flow and the production of biological organization, subsequently generalized in various ways…. Internal order can be produced by gradients of energy (matter/energy) flows through living systems. Structures so produced help not only draw more energy through the system, lengthen its retention time in the system, but also dissipate degraded energy, or entropy, to the environment, thus paying Schrödinger's "entropy debt." Living systems then are seen an instance of a more general phenomena of dissipative structures. [According to Jantsch] "[w]ith the help of this energy and matter exchange with the environment, the system maintains its inner non-equilibrium, and the non-equilibrium in turn maintains the exchange process…. A dissipative structure continuously renews itself and maintains a particular dynamic regime, a globally stable space-time structure"…. However, thermodynamics can deal only with the possibility that something can occur spontaneously; whether self-organizing phenomena occur depend upon the actual specific conditions (initial and boundary) as well as the relationships among components…

Seeing the cell as a thermodynamic "dissipative structure" was not to be considered as reducing the cell to physics, as Bernal pointed out, rather a richer physics of what Warren Weaver called "organized complexity" (in contrast to simple order or

> "disorganized complexity") was being deployed…. The development of this "new" physics of open systems and the dissipative structures that arise in them was the fulfillment of the development that Schrödinger foresaw…. Dissipative structures in physical and chemical systems are phenomena that are explained by nonequilibrium thermodynamics…. The emergent, self-organizing spatio-temporal patterns observed in the Belousov-Zhabotinski reaction are also seen in biological systems (such as in slime mold aggregation or electrical patterns in heart activity)… Indeed, related self-organizational phenomena pervade biology…. Such phenomena are seen not only in cells and organisms, but in ecosystems, which reinforces the notion that a broader systems perspective is needed as part of the new physics…. Important to such phenomena are the dynamics of non-linear interactions (where responses of a system can be much larger than the stimulus) and autocatalytic cycles (reaction sequences that are closed on themselves and in which a larger quantity of one or more starting materials is made through the processes). Given that the catalysts in biological systems are coded in the genes of the DNA, one place to start defining life is to view living systems as informed, autocatalytic cyclic entities that develop and evolve under the dual dictates of the second law of thermodynamics and of natural selection…. Such an approach non-reductively connects the phenomena of living systems with basic laws of physics and chemistry…. Others intuit that an even richer physics is needed to adequately capture the self-organizing phenomena observed in biology and speculate that a "fourth law" of thermodynamics about such phenomena may ultimately be needed…. In any event, increasingly the tools developed for the "sciences of complexity" and being deployed to develop better models of living systems…. Robert Rosen has reminded us that complexity is not life itself but what he terms "the habitat of life" and that we need to make our focus on the relational. "Organization inherently involves functions and their interrelations"…. Whether the existing sciences of complexity are sufficient or a newer conceptual framework is needed remains to be seen…. Living beings exhibit complex, functional organization and an ability to become more adapted to their environments over generational time, which phenomena represent the challenge to physically-based explanations based upon mechanistic (reductionistic) assumptions. By appealing to complex systems dynamics there is the possibility of physically-based theories that can robustly address phenomena of emergence without having recourse to the type of "vitalism" that was countenanced by some in the earlier part of the twentieth century.[144]

In other and fewer words, according to the dynamicist model of life, a living organism is essentially a far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic system, with special teleological dynamics (namely, reproduction, growth, motility, death, and evolution or natural selection), essential indexicality (namely, inherent context-dependency, together with egocentric centering in a frame of reference), and causal spontaneity (namely, efficacious metabolism, involving DNA, by means of epigenesis).

The contemporary Kantian dimension of this doctrine, in particular, raises a further important issue about how the biological and psychological properties of real human persons are cognized or known in the exact sciences, as Kant would have understood those sciences. For the purposes of correctly understanding Kant's conception, we must be able to see how it is no trivial fact that, in the 1750s, he wrote treatises on the rotation of the Earth, the age of the Earth, universal natural history, fire, earthquakes, and the theory of winds. Moreover, his

geography lectures were continuous with his anthropology lectures.[145] Thus Kant was in fact a *proto-theorist* of non-equilibrium thermodynamic systems, in a comprehensive sense, lacking only the essentially richer mathematics of complex systems dynamics and the other post-Kantian formal tools of modern logic, biology, chemistry, and physics, not to mention those of cognitive neuroscience and the social sciences.

Now Kant also had notoriously high standards for something's qualifying as a science. Not only must a science involve a systematic organization of objective facts or objective phenomena of some sort, it must also be *strongly nomological*, in the sense that it expresses necessary a priori laws (*MFNS* 4: 468). Sciences in this high-powered Kantian sense, in turn, can include either "constitutive" (that is, existentially committed without conditions, empirically meaningful or "objectively valid," and assertoric) principles or else "regulative" (that is, at best hypothetically existentially committed, logical-fictional or merely "thinkable," and non-assertoric) principles. An exact science can be a naturally mechanistic physical science—that is, an exact science whose subject-matter satisfies the conditions of Natural Mechanism—only if its phenomena and its laws are fully mathematically describable (*MFNS* 4: 470) in terms of recursive functions, which in turn are all Turing computable, according to The Church-Turing Thesis. But as I have argued elsewhere, Kant's notion of mathematics is significantly narrower than our contemporary notion.[146] So we must assume that full mathematical describability in terms of recursive functions *for Kant* is equivalent only to analyzability in terms of what is technically known as "primitive recursive arithmetic," because it is *the quantifier-free* theory of the natural numbers and the *primitive* recursive functions over the natural numbers—the successor function, addition, multiplication, exponentiation, etc.[147] Therefore for Kant, at least implicitly, a given theory will be a naturally mechanistic physical science only if its underlying mathematics is no more complex than primitive recursive arithmetic. Because primitive recursive arithmetic encodes all and only the primitive recursive functions, then obviously every function within its scope is also inherently Turing-computable.[148]

As we have seen, Kant regards biology as a merely regulative, non-mechanistic "life science" that supplements the classical Newtonian deterministic, mechanistic mathematical physics with the teleological concept of a natural purpose or living organism (*CPJ* 5: 369-415). But at the same time Kant regards this biological supplementation of physics as *explanatorily necessary*. And that is because biology for Kant provides representations of natural phenomena that are themselves explanatorily irreducible to deterministic mechanistic concepts. This thought is beautifully captured in a text I have quoted several times already:

> It is quite certain that we can never adequately come to know the organized beings and their internal possibility in accordance with merely mechanical principles of nature, let alone explain them; and this is indeed so certain that we can boldly say that it would be absurd for humans ever to make such an attempt or to hope that there might yet arise a Newton who could make comprehensible even the generation of a blade of grass according to natural laws. (*CPJ* 5: 400, underlining added)

In short, Kant has in effect anticipated the dynamicist model of life, with its leading conception of far-from-equilibrium, asymmetric, complex, self-organizing thermodynamics, and in so doing, he has in effect conceptually revolutionized the familiar classical notions of mechanistic causation and the linear equilibrium dynamics of inertial physical systems. Echoing the title of Mensch's book, we can think of the dynamicist model of life as implying

an *Organicist Revolution* that is fully comparable to Kant's "Copernican Revolution" in metaphysics. Kant's Copernican Revolution says that in order to explain rational human cognition and authentic a priori knowledge, we must hold that necessarily, the world structurally conforms to our minds, rather than the converse. The Organicist Revolution, in turn, says that the real possibility of human consciousness, cognition, caring, rationality, and free agency, and *therefore also the "Copernican" necessary structural conformity of world-to-mind, provided that we actually do exist*,[149] is built essentially into the non-equilibrium thermodynamics of organismic life, and necessarily underdetermined by naturally mechanical processes and facts. Hence the Organicist Revolution that is implied by the dynamicist model of life not only *includes* Kant's Copernican Revolution, but also goes one full revolutionary cycle *beyond* it.

One crucial further implication of the dynamicist model of life flows from the fact that the mathematics of non-equilibrium thermodynamics is essentially richer than primitive recursive arithmetic and Peano Arithmetic alike, in that it includes a full range of non-linear functions. Now Gödel's incompleteness theorems say

> (i) that there are logically unprovable true sentences in any elementary or classical second-order logical system that also includes enough axioms of Peano arithmetic, and
> (ii) that all such logical systems are consistent (that is, non-contradictory) if and only if they are incomplete (that is, not all the truths of the system are theorems of the system) and have their ground of truth outside the system itself.[150]

So Gödel's incompleteness theorems, taken together with The Church-Turing Thesis, jointly show that formal logical proof is not sufficient for mathematical truth, and also that mathematical truth itself is not a Turing-computable function that could be realized on a digital computing machine. Therefore mathematical truth itself, and especially including mathematical truths of non-equilibrium thermodynamics, are inherently uncomputable, non-naturally-mechanistic facts of physical nature.

The thesis of *ontological emergence* says that new, global or system-wide causally efficacious properties can arise in certain thermodynamic systems over time, and that these properties inherently change the overall thermodynamic constitution of the entire system.[151] This ontological emergence thesis is significantly metaphysically stronger than either the thesis of *epistemic emergence* (which merely says that thermodynamic systems can exemplify global relational properties that cannot be known or predicted by knowing the intrinsic non-relational properties of their parts together with their extrinsic law-governed modes of relational combination) or the minimal thesis of *historical emergence* (which merely says that thermodynamic systems can exemplify global relational properties at later times, that they did not exemplify at earlier times). Given the notion of a far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic system, then the dynamicist model of life predicts that there are natural systems of interacting material proper parts or elements whose actual behaviors over time can be neither digitally computed nor nomologically predicted due to random exchanges of causal information, energy, and matter with the surrounding environment, and that exemplify ontologically emergent, causally efficacious properties that are not strongly supervenient on the intrinsic non-relational properties of the elements of the system together with their extrinsic relational properties. These properties, as I have said, are

rich, causally efficacious immanent structures that dynamically emerge "between" the simpler pre-existing thermodynamic structures of equilibrium, temporally reversible, or otherwise mechanical systems, which ontologically "open up and unfold into" the far-from-equilibrium, temporally irreversible, complex dynamic system that thereby becomes causally dominant, just as the rich, continuous, nondenumerably infinite mathematical structures of the real number system constructively emerge—say, by means of the power set operation, or the Dedekind-cut operation—"between" the simpler, non-continuous, denumerably infinite mathematical structures of the pre-existing rational or natural number systems which mathematically "open up and unfold into" the more complex number system that thereby becomes quantitatively dominant.

This, again, is *dynamic emergence*. For example, according to the accounts provided by contemporary cosmological physics, it is plausibly arguable that The Big Bang (exemplifying thermodynamic expansion) and black holes (exemplifying thermodynamic collapse) are far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic systems with ontologically emergent properties.[152]

For my current purposes, what is most crucial is neither the non-trivial fact that The Big Bang and black holes are thermodynamic systems of this special kind, nor the equally non-trivial fact that the thesis of ontological emergence predicted by non-equilibrium thermodynamics is significantly more metaphysically robust than either a mere epistemic emergence thesis or a mere historical emergence thesis. Instead, what is most crucial is that, according to the dynamicist model of life, the conscious, caring, intentional, and rational biological and neurobiological processes of rational human animals, namely, real human persons, *also* constitute far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic systems and *also* exemplify ontological emergence. If so, then the real human personal activity of self-organization is inherently non-naturally-mechanistic in the strong sense that it inherently exceeds the reach of mere causal quantity-of-matter-and/or-energy conservation, together with Turing-computability, and also implements the abstract structures of naturally purposive or naturally teleological systems and intentional systems, while still being *fully compatible with*, although *necessarily underdetermined* by, all the general deterministic or indeterministic causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang.

The conscious, intentional, caring, and rational biological and neurobiological processes of real human persons are, as it were, and as I mentioned in the section 1.1, *Little Bangs*: small-scale causal singularities. Like all living organisms, they are really causally efficacious, because of their full compatibility with all the general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity of matter-and/or-energy facts about the past, especially including The Big Bang, yet their specific activities and operations are also *necessarily underdetermined* by this minimal cosmic matter-and/or-energy grid. So real human persons are, to that extent, *nomologically unique*. This means that via their conscious, intentional, caring, and rational living organismic, causally spontaneous choices and acts, they bring into existence "one-off," or one-time-only, causal-dynamical laws of rational human activity, that robustly enrich and supplement physical nature's repertoire of general causal laws.

According to this neo-Aristotelian, contemporary Kantian, dynamicist picture of physical nature, most explicitly—but unfortunately, also only fragmentarily—adumbrated by Kant himself in the *Opus postumum*, the complete set of deterministic-mechanical general causal

laws provides a minimal or skeletal causal-dynamic architecture for nature, which is then gradually elaborated, and creatively filled out, by the one-off laws of far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic systems. So according to the dynamicist model, not only does *every* thermodynamic process *increase* entropy, but also *some* thermodynamic processes also *dissipate* entropy, and thereby generate "negentropy," by increasing the structural ordering of nature, thereby actualizing its energetic potentialities in essentially novel ways, via natural purposiveness and natural teleology, especially by means of the causally spontaneous "productive capacity" of living organisms (*CPJ* 5: 421-425). In the case of organisms, this is the same as *epigenesis*. Through dissipative processes, especially those of organismic life via epigenesis, no new quantitites of matter and/or energy ever come into existence (that is, matter and/or energy conservation always obtains): only *new immanent structurings and structures of* matter and/or energy. A paradigmatic case of this is the intentional body-movements of minded animals, especially including rational human animals.[153]

As I pointed out in section 2.0, explanatory irreducibility is the irreducibility of certain mental representations to other mental representations, and by contrast, ontological irreducibility is the irreducibility of certain worldly properties to other worldly properties. Provided that there is a necessary one-to-one connection between distinct mental representations and distinct worldly properties, then explanatory irreducibility entails ontological irreducibility. Now Strong Kantian Non-Conceptualism guarantees the necessary one-to-one connection between the representation of life and the causally efficacious, immanent structural functional properties and/or facts that constitute living organisms. In this way, via Strong Kantian Non-Conceptualism, representational anti-mechanism yields dynamicism. Therefore, according to the neo-Aristotelian and contemporary Kantian dynamicist model of life, not only can there never be a *Newton* of a blade or grass, but also there can never be either a *Church* or a *Turing* of a blade of grass, or of the actual biological life of the rational human animal body, in both an explanatory sense and also an *ontological* sense, although Kant himself was more critically cautious about the latter.

Nevertheless, even for Kant's officially cautious view about the merely "regulative" and not "constitutive" semantic status[154] of teleological judgments and the concept of natural purposiveness alike, it remains true for Kant that necessarily, *were* it to be the case that natural purposes existed in the manifestly real natural world, then Natural Mechanism *would be* false, and also true that the specific modal character of the antecedent of that necessary subjunctive conditional is *real (synthetic) possibility*, not merely logical or conceptual (analytic) possibility. Moreover, even if Kant holds that *teleological judgments* are *not* constitutive, but instead only regulative, he can still consistently hold that we also have *constitutive teleological intuitions* of inner and outer sense, expressing pleasure and pain, aka "the feeling of life," that provide direct, veridical cognitive access to our own biological life, via the essentially non-conceptual content of essentially embodied experience. Hence on perfectly legitimate Kantian cognitive-semantic grounds, we can prove that at least some essentially non-mechanical, biological facts actually exist, *in us, as human animals*. So even Kant's officially cautious view is on the verge of committing itself to the ontological implications of my Kantian Non-Conceptualist version of representational anti-mechanism—the robustly anti-physicalist, anti-mechanist metaphysics of dynamicism.

We may now recall Chalmers's remark, quoted above as one of the epigraphs of this chapter, in strong support of physicalism about life, that is, in strong support of Natural Mechanism:

> Presented with a full physical account showing how physical processes perform the relevant functions, a reasonable vitalist would concede that life has been explained. There is not even *conceptual* room for the performance of these functions without life.

What I want to say in direct, three-part reply to Chalmers is this:

> (i) that a so-called "reasonable vitalist" is in fact *a Kantian Non-Conceptualist representational anti-mechanist*, and neither a substance vitalist nor a property or functional vitalist,
> (ii) that the relevant organismic functions are truly described by *the dynamicist model of life*, and neither by naturally mechanistic analysis, nor by "pop-out emergentist" functional analysis, and
> (iii) that even if "there is not even *conceptual* room for the performance of these functions without life," there is nevertheless more than enough *essentially non-conceptual room* for explanatorily and ontologically irreducible organismic life to perform its special far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic functions.

This sets the stage for the three anti-physicalist, anti-mechanist arguments I will spell out in the next and penultimate section of this chapter.

2.6 How Life Does Not Strongly Supervene on the Physical, and Why

As we saw in the last section, organismic life is not merely the occurrence of certain naturally mechanistic, Turing-computable behaviors, functions, or operations. More explicitly, organismic life is far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, thermodynamic activity with intrinsic teleological dynamics, essential indexicality, and causal spontaneity. If this is correct, then organisms occupy unique spatial locations in their environments, take unique paths through them when they are motile, and in any case necessarily include intrinsic spatiotemporal asymmetries. For example, every animal's body has right-left (top-bottom, front-back, etc.) asymmetries, and all metabolic processes are temporally irreversible processes. Moreover, as I argued above, it is plausible to hold that essential indexicality is the same as inherent context-dependency,[155] together with egocentric centering in a frame-of-reference, together with orientable space and irreversible time. Facts about self-organizing thermodynamic systems are therefore *essentially indexical* facts.

In *The Conscious Mind*, with his characteristic caution, Chalmers argues that indexical facts *might not* logically supervene on the fundamental physical facts:

> Does indexicality pose a problem for reductive explanation? For arbitrary speakers, perhaps not, as the "fact" in question can be relativized away. But for myself,

> it is not so easy. The indexical fact expresses something very salient about the world as I find it: that David Chalmers is *me*. How could one explain this seemingly brute fact? …. The issue is extraordinarily difficult to get a grip on, but it seems to me that even if the indexical is not an objective fact about the world, it is a fact about the world as I find it, and it is the world as I find it that needs explanation. The nature of the brute indexical is quite obscure, though, and it is most unclear how one might explain it…. The indexical fact may have to be taken as primitive. If so, then we have a failure of reductive explanation distinct from and analogous to the failure with consciousness.[156]

Here, Chalmers and I are in basic agreement. But I want also to go two steps further, and argue this:

> (i) that indexical facts *definitely do not* logically supervene on the physical, and
> (ii) that insofar as organismic life is an essentially indexical process, then it is not even *strongly* supervenient on the physical, much less logically supervenient.

Therefore physicalism about life, and correspondingly, Natural Mechanism about life, are both false.

Let us consider now the phenomenon of metabolism in living organisms, and the three following arguments.

> Argument 1: Inverted Life
>
> (1) Start with a representation of all the fundamental physical properties and facts about actual organismic metabolism. It is then representable, as a matter of conceivability, that actual organismic metabolism could be either (i) enantiomorphically reversed in space, or (ii) have its "time's arrow" systematically structurally deformed away from the classical time-model of continuous linear development (e.g., cyclical time, hyperbolic spiraling time, punctuated equilibrium time, etc.), while also representing that all the fundamental physical properties and facts in the world are held fixed.
> (2) We assume that Strong Kantian Non-Conceptualism is true with respect to the representation of life.
> (3) Therefore there is a real, objective instantiated or uninstantiated functional property "out there" in the world according to which, as a matter of conceivability, actual organismic metabolism could be either (i) enantiomorphically reversed in space, or (ii) have its "time's arrow" systematically structurally deformed away from the classical time-model of continuous linear development, while also all the fundamental physical properties and facts in the world are held fixed.
> (4) Therefore it is really possible that actual organismic metabolism could be either (i) enantiomorphically reversed in space, or (ii) have its "time's arrow" be systematically structurally deformed away from the classical time-model of continuous linear development, while also all the fundamental physical facts and properties in the world are held fixed.
> (5) Therefore the strong supervenience of biological life on the physical fails.

Argument 2: Suspended Life

(1) Start with a representation of all the fundamental physical properties and facts about actual organismic metabolism. It is then representable, as a matter of conceivability, that actual organismic metabolism could be universally frozen in actual time and actual place—that is, in a universal state of suspended animation without termination—while also representing that all the fundamental physical properties and facts in the world are held fixed.
(2) We assume that Strong Kantian Non-Conceptualism is true with respect to the representation of life.
(3) Therefore there is a real, objective instantiated or uninstantiated functional property "out there" in the world according to which, as a matter of conceivability, actual organismic metabolism could be in a universal state of suspended animation without termination, while also all the fundamental physical properties and facts in the world are held fixed.
(4) Therefore it is really possible that actual organismic metabolism could be in a universal state of suspended animation without termination, while also all the fundamental physical facts and properties in the world are held fixed.
(5) Therefore the strong supervenience of biological life on the physical fails.

Argument 3: Non-Local Life

(1) Start with a representation of all the fundamental physical properties and facts about actual organismic metabolism. It is then representable, as a matter of conceivability, that actual organismic metabolism could be spread over the universe in such a way that it lacks unique location and causal determinacy—as in non-locality and indeterminacy effects in quantum mechanics, for example, Schrödinger's cat paradox[157]—while also representing that all the fundamental physical properties and facts in the world are held fixed.
(2) We assume that Strong Kantian Non-Conceptualism is true with respect to the representation of life.
(3) Therefore there is a real, objective instantiated or uninstantiated functional property "out there" in the world according to which, as a matter of mere conceivability, actual organismic metabolism could be spread over the universe in such a way that it lacks unique location and causal determinacy, while also all fundamental physical properties and facts in the world are held fixed.
(4) Therefore it is really possible that actual organismic metabolism could be spread over the universe in such a way that it lacks unique location and causal determinacy, while also all fundamental physical properties and facts in the world are held fixed.
(5) Therefore the strong supervenience of biological life on the physical fails.

These three arguments, respectively, are relevantly analogous to Chalmers's formulations of

(i) *the inverted qualia argument* for the irreducibility of phenomenal consciousness, which entails the failure of the strict or logical determination of the specific character of phenomenal consciousness by the physical,

> (ii) *the zombie argument* for the non-reducibility of phenomenal consciousness, which entails the failure of the strict or logical determination of the existence of phenomenal consciousness by the physical, and
> (iii) *the pan(proto)psychist argument* for the possibility of universal proto-mentality in a physical world, which shows that some version of *neutral monism* is possible.

So if Chalmers's three arguments are (arguably) sound, then, by the same token, so are mine.

There is a very big, further difference between my arguments and Chalmers's, however. This extra mega-difference lies in the fact that although, like Chalmers, I have grounded the inferential step to possibility on *conceivability*, nevertheless, because I am also assuming the truth of immanent structuralism about properties and also of Strong Kantian Non-Conceptualism, it is thereby guaranteed that the irreducible essentially non-conceptual representation of life picks out real, objective, *causally efficacious* immanent structural functional properties of living organisms. Thus the real possibilities I have identified *are sufficient to undermine the strong supervenience of organismic life on the physical, not merely its logical supervenience*. By sharp contrast, Chalmers's irreducibility arguments leave nomological or natural strong supervenience in place, and thereby pick out only *epiphenomenal* mental properties that are characteristic of either substance dualism or property dualism, metaphysically floating beyond or above the fundamental physical world.[158] The epiphenomenalism problem, as Jaegwon Kim has repeatedly pointed out, is that all classically dualistic properties are causal-explanatorily *excluded* by the fundamental physical properties on which those dualistic properties nomologically or naturally strongly supervene, and as a consequence those classically dualistic properties are rendered causally and explanatorily *inert*.[159]

By sharp contrast, according to representational anti-mechanism and the dynamicist model of life, in a Kantian Non-Conceptualist framework, the essentially non-conceptual representation of life picks out a set of objective, real immanent structural thermodynamic properties of living organisms, and these living organisms are all causally efficacious. Indeed, in a sense, reductive physicalists about life completely *agree* with my thesis that biological life is causally efficacious, which is precisely why they attempt to reduce all the causal powers of organisms to the causal powers of fundamental physical properties, lest living systems be rendered causally inert. They merely disagree with my thesis that biological life, *as* organismic life, is necessarily underdetermined by and irreducible to fundamental physical properties.

Nevertheless, above all, as per the arguments from Inverted Life, Suspended Life, and Non-Local Life, the causally efficacious, objective, real fact of organismic life does *not* strongly or logically supervene on the Turing-computable deterministic or indeterministic causal behaviors, functions, operations, and/or states bound up with fundamental physical properties and facts. So Natural Mechanism is false. Correspondingly, and ironically, Chalmers is "dead wrong" about the reducibility of life to the physical, even though he is absolutely right about the actual irreducibility of consciousness and the possible irreducibility of indexicality—although not for the specific reasons he gave.

Essentially the same basic anti-mechanist philosophical points I have just made were also made by Hans Jonas in the mid-1960s:

> Suppose that it is a living body, an organism, on which the gaze of the divine mathematician happens to rest. It may be unicellular or multicellular. What would the God of the physicists "see"? As a physical body the organism will exhibit the same general features as do other aggregates: a void mostly, crisscrossed by the geometry of forces that emanate from the insular foci of localized elementary being. But special goings-on will be discernible, both inside and outside its so-called boundary, which will render its phenomenal unity still more problematical than that of ordinary bodies, and will efface almost entirely its material identity through time. I refer to its *metabolism*, its exchange of matter with the surroundings. In this remarkable mode of being, the material parts of which the organism consists at any moment are to the penetrating observer only temporary, passing contents whose joint material identity does not coincide with the identity of the whole which they enter and leave, and which sustains its own identity by the very act of foreign matter passing through its spatial system, the living *form*…. [T]he object-view of the divine mathematician is less concrete and colorful than ours—but would we also grant it, as before, the possibility of being truer? Emphatically not in this case, and here we move on firm ground, because here, being living bodies ourselves, we happen to have inside knowledge. On the strength of the immediate testimony of our bodies *we* are able to say what no disembodied on looker would have a cause for saying: the mathematical God in his homogenous analytical view misses the decisive point—the point of life itself: its being self-centered individuality, being for itself and in contraposition to all the rest of the world, with an essential boundary dividing "inside" and "outside"—notwithstanding, nay, on the very basis of the actual exchange.[160]

To be sure, Jonas formulates his points within the framework of existential phenomenology, and not (or at least not explicitly) within my favored and more inclusive four-part framework of

> (i) the dynamicist model of life, together with
> (ii) immanent structuralism,
> (iii) Strong Kantian Non-Conceptualism, and
> (iv) the three non-reductive arguments from the conceivability of Inverted Life, Suspended Life, and Non-Local Life.

But I do also think that this more inclusive four-part framework is fully receptive to the basic ideas of Jonas's approach.

In any case, in this chapter I have worked out and defended a neo-Aristotelian and contemporary Kantian anti-mechanist, anti-physicalist, anti-dualist, dynamicist philosophy of biology, which says that, in view of non-equilibrium/complex system thermodynamics, the necessary and sufficient conditions of the real possibility of biological life are:

> (i) *intrinsic teleological dynamics in organisms*: their self-organizing intentionality, including reproduction, growth, motility, death, and evolution or natural selection,
> (ii) *essential indexicality in organisms*: their inherent context-dependency, together with egocentric (although not necessarily conscious) centering in a frame of reference, together with orientable space and irreversible time, and

> (iii) *causal spontaneity in organisms*: their efficacious metabolism, involving DNA, by means of epigenesis.

So, in effect, I am proposing a *real-metaphysical definition*[161] of biological life, using the dynamicist model of life and the other three elements of the four-part framework I have proposed.

In this connection, however, it needs to be emphasized that I am fully aware that for many or most contemporary working scientists and philosophers of biology, there is no general consensus about the definition of life—although that could be finessed away as the normal professional academic condition of "scholarly disagreement"—and also, more importantly, that there are some of these who hold that, even in principle, no such thing as a philosophical definition of life is possible,

> either (i) because the concept LIFE is merely a family-resemblance concept of some sort, with no inherent epistemic or semantic unity,
> or (ii) because the concept LIFE is purely instrumental or pragmatic, and strongly relative to past, contemporary, or future scientific practices,
> or (iii) because the very idea of a priori philosophical conceptual definitions, per se, is epistemically or metaphysically suspect.[162]

On my view, however, the representation of life is not a *concept* at all, but instead an essentially non-conceptual content. Hence all of these worries about the concept LIFE, or conceptual definitions, even if they are correct, are strictly beside my main point. Indeed, in order for any of these worries to be directly relevant to my argument, a convincing demonstration of the truth of Conceptualism would already have to be in hand. But as I have already argued in detail and at length elsewhere, Conceptualism is false.[163]

Finally, it needs to be *re*-reemphasized that this neo-Aristotelian and contemporary Kantian anti-mechanist, anti-physicalist, anti-dualist, dynamicist philosophy of biology is *neither* a version of substance vitalism, requiring an appeal to vital spirit or stuff of some sort, *nor* is it a version of property or functional vitalism, requiring an appeal to the nomologically supervenient, synchronic, static, "pop-out" emergence of essentially distinct vital properties and/or facts. On the contrary, it is entirely a real metaphysics of the immanent structures of non-equilibrium thermodynamic systems, and entails at most the non-strongly-supervenient, diachronic, dynamic emergence of certain inherently non-causally-nomologically-determined, uncomputable, necessary a priori, immanent structural dynamicist properties in living organisms, in a way that is also perfectly consistent with all the deterministic and indeterministic causal laws of nature, especially including the Conservation Laws. What is essential is simply that these dynamicist properties are *not necessarily determined, or entailed, by these laws*, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang. As a consequence, I think that my view fully satisfies the scientifically-robust methodological constraints on teleological biological explanations famously spelled out by Ernest Nagel in "Teleology Revisited,"[164] while at the same time *avoiding* the errors of Natural Mechanism.

2.7 CONCLUSION

For all the reasons I have given in this chapter, the dynamicist model of life, when taken together with a neo-Aristotelian and contemporary Kantian theory of immanent structural properties, representational anti-mechanism, and Strong Kantian Non-Conceptualism as applied to the representation of life, jointly provide the basis of a very powerful cognitive-semantic and metaphysical argument for explanatory and ontological anti-physicalism which, if it is sound, shows that the causally efficacious fact of biological life, *as* organismic life, does not strongly supervene in any way on the fundamental physical world, and correspondingly, shows that not only is Natural Mechanism is false, but also any version of anti-physicalist, anti-mechanist dualism about life is also false.

Above all, however, *organisms* are not machines, *animals* are not machines, *minded animals* are not machines, and *you* are not a machine. So LaMettrie,[165] the ultra-Darwinian biologist Thomas Huxley,[166] Sam Harris,[167] Daniel Dennett,[168] and scads of less-well-known contemporary philosophers, too numerous to mention, are all "dead wrong" about that too.

Nor are you a ghost, with essentially mysterious causal powers. So Descartes and defenders of Classical Agent-Causal Libertarianism are equally "dead wrong" about *that*, too.

This leads to one last important thing in this connection—perhaps even the most important thing, at the end of the philosophical day. This neo-Aristotelian and contemporary Kantian anti-mechanist, dynamicist approach to biology and the phenomenon of life can significantly inflect or even radically modify our philosophical concept or picture of physical nature itself in a deeply non-Newtonian and non-Cartesian way, by representing the causal behaviors, functions, operations, and/or states of fundamental physical properties and facts *themselves* as inherently open to the real possibility of organismic life, consciousness, intentionality, caring, deeply free agency, and real personhood.[169] In turn, as the later Wittgenstein and Jonas both forcefully remind us, a change in form of representation is also a change in our cognitive and practical *attitudes*, which in turn directly affect or even motivate intentional action.[170] As human *animals*, we are ineluctably embedded in physical nature. So to change our philosophical concept or picture of physical nature along the liberalized lines proposed in this chapter is not only to change our cognitive and practical attitudes *towards the natural universe*, but also to change our cognitive and practical attitudes *towards our own lives*. That is the ultimate upshot of the *natural piety* approach to the philosophy of nature and natural science.[171]

According to natural piety, *neither* are you alienated from nature (a Cartesian ghost-in-a-machine) *nor* are you a "lord and master" of nature (a Baconian/Cartesian technocrat). To believe both of these at once was Victor Frankenstein's tragic mistake, repeated endlessly and magnified infinitely in the deeply misguided epistemic and metaphysical doctrines, and scientistic-technocratic ideology, of Natural Mechanism:[172]

> Learn from me, if not by my precepts, at least by my example, how dangerous is the acquirement of [naturally mechanistic] knowledge, and how much happier that man is who believes his native town to be the world, than he who aspires to become greater than his nature will allow.[173]

Or—jumping forward from early 19th century European intellectual culture to late 20th century European intellectual culture—as Prigogine puts it:

> The attempt to understand nature remains one of the basic objectives of Western thought. It should not, however, be identified with the idea of control. The master who believes he understands his slaves because they obey his orders would be blind. When we turn to physics, our expectations are obviously different, but here as well, Vladimir Nabokov's conviction rings true: "What can be controlled is never completely real; what is real can never be completely controlled." The [naturally mechanistic] classical ideal of science, a world without time, memory, and history, recalls the totalitarian nightmares described by Aldous Huxley, Milan Kundera, and George Orwell.[174]

As against Natural Mechanism, according to natural piety, the physical universe as a whole and in all its parts, in its basic mathematical and thermodynamic structures, and especially the localized versions of all these structures that are realized on planet Earth, is "our town."[175] And in "our town," in *this* actual physical universe and on *this* very planet, you just *are* your minded animal life, as I will argue in chapters 6-7 below. Therefore you are morally, metaphysically, and epistemically *at home* in nature, for better or worse.

Chapter 3

FROM BIOLOGY TO AGENCY

Freedom in the practical sense is the independence of the power of choice (*Willkür*) from **necessitation** by impulses of sensibility. For a power of choice is **sensible** insofar as it is pathologically affected (through moving-causes of sensibility); it is called an animal power of choice (*arbitrium brutum*) if it can be **pathologically necessitated**. The human power of choice is indeed an *arbitrium sensitivum*, yet not *brutum*, but *liberum*, because sensibility does not render its action necessary, but in the human being there is a faculty of determining oneself from oneself, independently of necessitation by sensible impulses. (*CPR* A534/B562, boldfacing in the original)

Practical freedom can be proved through experience. For it is not merely that which stimulates the senses, i.e., immediate affects them, that determines human choice, but we always have a capacity to overcome impressions on our sensory faculty of desire by representations of that which is useful or injurious even in a more remote way; but these considerations about that which in regard to our whole condition is desirable, i.e., good and useful, depend on reason. Hence this also yields laws that are imperatives, i.e., objective **laws of freedom**, and that say **what ought to happen**, even though it never does happen…. We thus cognize practical freedom through experience, as one of the natural causes, namely

a causality of reason in the determination of the will. (*CPR* A802-803/B830-831, boldfacing in the original)

THE HUMAN BEING AS A BEING IN THE WORLD, SELF-LIMITED THROUGH NATURE AND DUTY. (*OP* 21: 34)

It is only because a person has volitions of the second order that he is capable both of enjoying and lacking freedom of the will.[176]

Kant is sometimes thought of as a cold, dry, rationalist. But he is really an emotional extremist.[177]

3.0 INTRODUCTION

In this chapter, building on and extending the neo-Aristotelian and contemporary Kantian dynamicist philosophy of biology that I worked out and defended in chapter 2, I will work out and defend a corresponding contemporary Kantian theory of practical freedom.

"Practical freedom" in Kant's terminology is what I have been calling *free agency* (= free will + practical agency). Free agency in this sense presupposes, but also significantly exceeds, what Kant calls "transcendental freedom," or what I have been calling *deep freedom*, *ultimate sourcehood*, or *up-to-me-ness*. More precisely, however, free agency also includes the capacity for what Kant calls "autonomy," or rational self-legislation, as well as the capacity for what the Existentialists, and more recently Harry Frankfurt,[178] have called "authenticity," "purity of heart," or "wholeheartedness." The fusion of the capacities for autonomy in the Kantian sense and for purity of heart or wholeheartedness is what I will call the complex capacity for *principled authenticity*. I think that this contemporary Kantian conception of free agency, as the metaphysical combination of our complex capacities for deep freedom and principled authenticity, is not only fully intelligible, and philosophically liberating—by virtue of its anti-mechanist/natural- pietist and non-intellectualist foundations, and its heavy-duty downstream ethical and political implications (see especially sections 3.2 to 3.3, as following on from sections 2.2 to 2.5)—but also objectively *true*, and I will argue for it as such.

3.1 TWO-DIMENSIONAL RATIONAL NORMATIVITY

In this chapter and throughout the rest of the book, I will presuppose a certain basic background conception of normativity that I call *The Two-Dimensional Conception of Rational Normativity*, aka The 2D Conception, for the purposes of the real metaphysics of free will and practical agency, but whose specific application to the philosophy of mind and knowledge I have also worked out in *Cognition, Content, and the A Priori*, section 1.2.

What is rational normativity? As I am using this notion, rational normativity is the following irreducible two-part fact

> (i) that all rational animals or real persons have aims, commitments, ends, goals, ideals, and values (hence, as rational animals, they are also *teleological animals*), and
> (ii) that these rational animals or real persons naturally treat their aims, commitments, ends, goals, ideals, and values (hence, as rational and teleological animals, they naturally treat these telic targets), (iia) as rules or principles for guiding theoretical inquiry and practical enterprises, (iib) as reasons for justifying beliefs and intentional actions, and also (iic) as standards for critical evaluation and judgment.

Furthermore, rational normativity in this sense can be

> either (i) *instrumental*, that is, conditional, hypothetical, desired for the sake of some further desired end, pragmatic, prudential, or consequence-based,

> or (ii) *non-instrumental*, that is, unconditional, categorical, desired for its own sake as an end-in-itself, non-pragmatic, non-prudential, and obtain no-matter-what-the-consequences.

As such, norms provide reasons for belief, cognition, knowledge, and intentional action; and categorical norms provide *overriding* reasons for belief, cognition, knowledge, and intentional action. Categorical norms are perfectly consistent with norms that are instrumental, conditional, desired for the sake of other ends, pragmatic, prudential, or obtain only in virtue of good consequences. Nevertheless, categorical norms are *necessarily underdetermined* by all other sorts of norms, and therefore cannot be assimilated to or replaced by those other sorts of norms.

Granting this general teleological conception of rational normativity and rational norms as a theoretical backdrop, then as I see it, there are two importantly distinct *kinds* of rational normative standards:

> (i) *minimal or nonideal standards*, which specify a "low-bar" set of goals, targets, principles, or rules, below which normatively evaluable activity cannot and does not occur at all, and which therefore jointly constitute a qualifying level of normativity, and
> (ii) *maximal or ideal standards*, which necessarily include and presuppose the (satisfaction of the) minimal, non-ideal, or low-bar standards, but also specify a further "high-bar" set of goals, targets, principles, or rules, below which normatively evaluable activity indeed occurs, but is always more or less imperfect, and in certain relevant respects, bad activity, and above which more or less perfected, and in the relevant respects, fully good activity occurs, and which therefore jointly constitute a perfectionist level of normativity.

So if I am correct, then

> (i) all rational normativity includes both low-bar or qualifying standards and also high-bar or perfectionist standards,
> (ii) the satisfaction of the high-bar standards necessarily requires the satisfaction of the low-bar standards,
> (iii) the satisfaction of the low-bar standards is not in itself sufficient for the satisfaction of the high-bar standards, but also
> (iv) failing to satisfy the high-bar standards is not in itself sufficient for failing to satisfy the low-bar standards.

This, in a nutshell, is The 2D Conception.

One fundamental implication of The 2D Conception, captured in clause (iv), is that it opens up the philosophically and evaluatively important possibility that intentional activities, or persons, can fail to meet the maximal, ideal, or high-bar standards without thereby failing to meet the minimal, nonideal, or low-bar standards. In this way, correspondingly, intentional activities, or persons, can fail to meet the high-bar standards without thereby falling off the scale of normative evaluability, and without ruling out appropriate attribution or acceptance of responsibility, which in turn go necessarily together with the possibility of rational learning and improvement. In other words, you can fall substantially short of being perfect, without also

getting off the hook of being rationally obligated to at least trying to do it significantly better, or to at least trying to be a significantly better person, the next time around.

Here is another fundamental implication of The 2D Conception. Just because someone fails to be maximally or ideally logically rational, by occasionally or even quite often failing to have correct logical insights, or by occasionally or even quite often committing logical fallacies in theoretical reasoning, it does not follow that this person is logically irrational, provided that she meets the minimal, nonideal, or low-bar standards of possessing basic logical knowing capacities and basic logical reasoning capacities. These basic capacities can, if correctly used, be perfected and realized, under appropriate other-things-being-equal conditions, and then at least *sometimes* meet at least some of the maximal, ideal, or high-bar logical reasoning standards, such as knowing logical truths, and carrying out consistent, valid, and sound inferences.

Or in other words, given The 2D Conception, even if a great many or even most people very often or even mostly are lacking in logical insight, and frequently commit logical fallacies in theoretical reasoning, nevertheless it does not follow that "people are logically irrational animals," provided that they still possess online logical knowing and logical reasoning capacities.[179] If this is so, then they minimally qualify for acquiring logical knowledge and for carrying out logical reasoning activity and for normative evaluability in this crucial domain of their lives as rational human animals. In the real world and on the ground, to be sure, as "human, all too human" animals, they may on the whole be quite unsuccessful at acquiring logical knowledge and carrying out logical reasoning. But at the same time, they are still to some extent rationally responsible for their own logical failings and errors, and in direct proportion to that level of responsibility, still capable of logical learning and logical improvement. Their logico-cognitive situation is far from perfect, but neither is it rationally hopeless, nor is it really impossible.

Correspondingly, and even more importantly, just because someone fails to be maximally or ideally morally or practically rational, by occasionally or even quite often choosing and acting immorally, and by occasionally or even quite often committing mistakes in practical reasoning, it does not follow that this person is morally or practically irrational, provided that she meets the minimal, nonideal, or low-bar standards of possessing the basic online capacity for free will, and basic moral/practical reasoning capacities. These basic capacities can, if correctly used, be perfected and realized, under appropriate other-things-being-equal conditions, and then at least sometimes meet at least *some* of the maximal, ideal, or high-bar moral rationality standards, such as at least occasionally choosing and doing the right things for the right reasons. Or in other words, given The 2D Conception, even if a great many or even most people very often or even mostly make wicked choices and do wicked things, and frequently make serious mistakes in moral or practical reasoning, it does not follow that "people are morally or practically irrational animals," provided that they still possess online free volitional and moral/practical reasoning capacities. If this is so, then they minimally qualify for pursuing morally right choice and action, and for carrying out good practical reasoning activity, and for normative evaluability in this crucial domain of their lives as rational human animals. In the real world and on the ground, to be sure, as "human, all too human" animals, they may on the whole be quite unsuccessful in their pursuits of right choice or action and at carrying out good moral reasoning. But at the same time they are still to some extent rationally responsible for their own errors, and in direct proportion to that level of rational responsibility, also capable

of moral learning and improvement. Their moral-practical situation is far from perfect, but neither is it rationally hopeless nor is it really impossible.

More generally, then, the all-too-common, facile, and overhasty "inference-to-human-irrationality"—whether in the form of logical, moral, or practical irrationality, or in the form of some other supposed basic incapacity of human animals per se, for example, epistemic irrationality, emotional irrationality, etc.—from the actual fact of more or less widespread error, or of more or less widespread imperfectly conducted activity, is itself simply an informal fallacy, and, ironically enough, therefore a "logical sin"[180] that the fallacious inference to human irrationality will not itself excuse, based essentially on a failure to recognize the fundamental and synoptic character of The 2D Conception.

These points, in turn, make it possible to see very clearly the fundamental flaw in what I will call *The One-Dimensional Conception of Rational Normativity*, aka The 1D Conception, no matter how plausible and sophisticated the theories that fall under this rubric might otherwise be. According to The 1D Conception, any failure to meet the maximal, ideal, or high-bar standards of rational normativity, or at the very least any failure to try to meet the high-bar standards, entails non-rationality, non-agency, and non-responsibility. Notice here, too, that the "trying" that is required or obligatory according to The 2D Conception, is just endeavoring to learn and improve one's logical, moral, practical, etc., performance relative to the perfectionist standard, and relative to the normative "distance" between the minimal/nonideal and maximal/ideal standards.

According to The 2D Conception, then, you are not rationally obligated even to *try* to *be* perfect, because perfection is always supererogatory, no matter how wonderful it might be—which may come as somewhat of a relief to you. You are rationally obligated only to try to learn and to improve your logical, moral, practical, etc., performance relative to the perfectionist standard, and relative to the normative "distance" between the minimal/nonideal and maximal/ideal standards. That is, according to the sane perfectionism included in The 2D Conception, you are rationally obligated only to try to ameliorate your imperfections, and better yourself by significant improvement, not rationally obligated either to try to become perfect, or to be perfect. Perfection, although a maximal or ideal rational-agent-centered normative standard, nevertheless exceeds what is rationally obligatory for rational, "human, all too human" animals like us. In Kantian language, perfection is a *regulative* standard, not a *constitutive* standard (*CPR* A642-668/670-696). Trying to change your life for *the better* is fully good enough: you don't have to (even try to be) be *the best*. The alternative is an insane perfectionism that is, in effect, covertly designed to guarantee failure for most agents in most contexts of agency in this thoroughly nonideal world. Thus to extend Voltaire's famous aphorism, "the best is the enemy of the good,"[181] according to the ameliorative (im)perfectionism I am proposing on the basis of the 2D Conception, *the better is the good's best friend*.

In any case, as an excellent contemporary defender of The 1D Conception, in *Self-Constitution*, Christine Korsgaard writes:

> I am eventually going to argue that bad actions, *defective* actions, are ones that fail to constitute their agents as the unified authors of their actions.
>
> [T]he hypothetical and categorical imperatives are constitutive principles of volition and action. Unless we conform to them—unless we are at least trying to conform to them—we are not willing or acting at all.

> The *Groundwork* portrays bad action as heteronomous action. Commentators often complain that if that is supposed to mean action that is caused by external forces, it is impossible to see how people are ever responsible for bad action. But of course the problem is much deeper than that, for if a person's movements are caused by external forces, it is not clear why we should call them actions at all.
>
> It is the essential nature of action that it has a certain metaphysical property—autonomy in Kant's argument, constitutional unity in Plato's. This explains why action must meet the normative standard: *it just isn't action* if it doesn't. But it also seem as of it explains it rather too well, for it seems to imply that only good action really is action, and that there is nothing left for bad action to be. [182]

Or in other words, according to Korsgaard, if you are not meeting or not trying to meet the classical perfectionist standards of human rationality, then you are a rationally defective and irrational animal, and off the hook. For example, if you fail to reason or fail to try to reason in a classical-logic-perfectionist way—that is, if you fail to conform fully to all the semantic or inferential standards of, say, classical sentential logic, or classical first-order predicate logic—then you are not in any sense a rational or responsible logical agent. Or if you fail to choose or act, or fail to try to choose or act, in a morally or practically classically perfectionist way—for example, by having a *good will* in Kant's sense (*GMM* 4:393) (*CPrR* 5: 110), by occurrently partially or completely realizing autonomy—then you are not in any sense a rational or responsible moral or practical agent. Disastrously, these results of One-Dimensionalism play directly into the hands of radical logical, moral, practical, and other sorts of skeptics, since as a matter of fact no actual rational human animal ever manages to meet or try to meet all or even most of the high-bar standards of rational normativity, but instead is doing extremely well, and indeed is doing something supererogatory and morally heroic, if she ever manages to meet or try to meet even some of them. How convenient for the radical skeptic!, then, that most or all of us, most or all of the time, turn out to be irrational animals.

Perhaps even more disastrously, these untoward results of One-Dimensionalism also play directly into the hands of individual "human, all too human" intentional agents looking for a fast track out of their everyday logical, moral, practical, etc., difficulties in a thoroughly nonideal world. How convenient for me, sinner that I am, that falling far short of rational perfection or falling far short of trying for rational perfection should entail the suspension of my responsibility. —If 1D Rational Normativity is Dead, then Everything is Permitted, and in turn I can simply take *the nihilist's way out* and then do the wrong thing without compunction, like the pathetically wicked character Smerdyakov in *The Brothers Karamazov*:

> "Take that money away with you, sir," Smerdyakov said with a sigh.
>
> "Of course, I'll take it! But why are you giving it to me if you committed a murder to get it?" Ivan asked, looking at him with intense surprise.
>
> "I don't want it at all," Smerdyakov said in a shaking voice, with a wave of the hand.
>
> "I did have an idea of starting a new life in Moscow, but that was just a dream, sir, and mostly because 'everything is permitted'. This you did teach me, sir, for you talked to me a lot about such things: for if there's no everlasting God [that is, in effect, 1D Rational Normativity + Divine Command Ethics], there's no such thing as virtue, and there's no need of it at all. Yes, sir, you were right about that. That's the way I reasoned."[183]

For all these reasons, it is clear that The 1D Conception of Rational Normativity is false, and also plausibly arguable that The 2D Conception is true.

3.2 KANT'S BIOLOGICAL THEORY OF FREEDOM

Kant was the first post-Newtonian philosopher to attempt to face up directly and fully to the basic philosophical problems of free will and Universal Natural Determinism. Prior to the 18th century, philosophers had always addressed issues about free will in the context of either classical Fatalism or Universal Divine Determinism—that is, in the context of "God-soaked" metaphysics, or what Kant calls *transcendental theology*. But as Kant so brilliantly saw, by the middle of the 18th century the philosophical context for thinking about free will had decisively shifted from transcendental theology to deterministic, mechanistic natural science, namely, to Newtonian mathematical physics. In that new context, other 18th century post-Newtonian philosophers, like Hume, focused almost exclusively on trying to provide a phenomenology or philosophical psychology of free will, as opposed to a metaphysics of free will.[184] But neither pre-18th century philosophers nor other 18th century post-Newtonian philosophers had clearly framed the free will problem as a puzzle *both* about explaining the possibility of free will in a universally determined, essentially mechanical natural world and *also* about the compatibility or incompatibility of free will and Universal Natural Determinism. That was one of Kant's most brilliant contributions to the philosophy of free will.

Otherwise put, Kant was the first post-Newtonian philosopher to see clearly and distinctly that Incompatibilism is consistent with Hard Determinism and Classical Libertarianism *alike*, both of which, in turn, are mistakenly committed to a deeper philosophical mistake. This deeper mistake, according to Kant, is the failure to distinguish metaphysically and epistemically between (in effect, veridical) appearances, aka phenomena, and things-in-themselves, aka noumena. And that failure leads directly to a logico-metaphysical antinomy, the Third Antinomy of Pure Reason. But above all, this deeper mistake inevitably hides the real philosophical significance of Compatibilism. In chapters 4 and 5 below, I will carefully lay out and critically discuss the basic competing positions in the free will debate, including Divine Fatalism, Universal Divine Determinism, Hard Determinism, Classical Libertarianism (including its agent-causal, non-causal, and event-causal indeterminist versions), and various versions of Incompatibilism and Compatibilism. The crucial point in the present context is just that Kant was brilliantly unique by trying to address *both* the metaphysics and the phenomenology/philosophical psychology of free will on the one hand, and *also* the classical Compatibilism vs. classical Incompatibilism dilemma on the other, within the new non-theological, post-Newtonian context of mechanistic Universal Natural Determinism.

In the rest of this section, I will focus on explaining and defending a new interpretation of Kant's theory of what he calls "transcendental freedom." Kant's theory of transcendental freedom is his metaphysics of free will. Transcendental freedom is how a rational animal or person can, "from itself" (*von selbst*) (*CPR* A533/B561), be the spontaneous mental cause of certain natural events or processes. If I am that rational animal or person, then insofar as I am transcendentally free, it follows that certain events or processes in physical nature are *up to*

me—or to use Kant's own phrase, *in meiner Gewalt*, literally: "in my control" or "in my power" (*CPrR* 5: 94-95). So otherwise put, transcendental freedom is deep freedom of the will, up-to-me-ness, or as it were (since it does not quite scan in grammatically correct German), *In-Meiner-Gewalt-Sein*.

In this connection I will argue, contrary to standard interpretations,[185] that Kant's theory of transcendental freedom entails neither a classically incompatibilist Timeless Agency theory nor a classically compatibilist Regulative Idea theory, and also that it thereby constitutes what Hodgson, in a text I used as an epigraph for chapter 1, aptly calls a "third alternative" to the all-too-familiar and seemingly exhaustive dichotomy between Incompatibilism and Compatibilism. This third alternative is a version of Incompatibilistic Compatibilism that I call *Kant's Biological Theory of Freedom*.

Now practical freedom presupposes transcendental freedom, and can be defined in a negative way as the independence of first-order volition, or the "power of choice" (*Willkür*), from necessitation by sensible impulses (*CPR* A533/B561). But practical freedom is also necessarily equivalent to what Kant calls "autonomy":

> the moral law expresses nothing other than the *autonomy* of pure practical reason, that is, [practical] freedom (*CPrR* 5: 33).

Practical freedom or autonomy is how a transcendentally free person can choose or do things by means of her subjective experience, or consciousness, of recognizing the Categorical Imperative or moral law as a desire-overriding, strictly universal, a priori, categorically normative, non-instrumental practical reason that has both motivating and justifying force. In turn, the special fact of this subjective experience, or consciousness, of autonomous agency is what Kant calls the "fact of reason" (*Faktum der Vernunft*):

> The consciousness of this fundamental law [of pure practical reason, which says: so act that the maxim of your will could always hold at the same time as a principle of universal law giving] may be called a fact of reason, since one cannot ferret it out from antecedent data of reason, *such as the consciousness of freedom* (for this is not antecedently given), and since it forces itself upon us as a synthetic proposition a priori based on no pure or empirical intuition… In order to regard this law without any misinterpretation as given, one must note that it is not an empirical fact, but the sole fact of pure reason, which by it proclaims itself as originating law. (*CPrR* 5: 31, underlining added).

So otherwise put, practical freedom or autonomy is *rational causation*. In section 3.3, I will argue, again contrary to standard interpretations, that a Kantian theory of practical freedom or autonomy entails a special form of *internalism about practical reasons* that shares something important with Hume's theory of practical reasoning. More specifically, I will focus on explaining and defending a *contemporary Kantian* theory of practical freedom or free agency, against the backdrop of *Kant's* Biological Theory of Freedom.

The basic link between the topics of the two sections—thus the basic link between Kant's Biological Theory of Freedom and the contemporary Kantian theory of practical freedom/free agency that I will develop—is Kant's theory of *teleology*, namely, his theory of *value-laden ends or purposes, and goal-directed activity*. So in the rest of this section, I will spell out and

philosophically exploit Kant's theory of *natural* teleology in the two Introductions and second half of the *Critique of the Power of Judgment*; and then in section 3.3, correspondingly, I will spell out and philosophically exploit his theory of *moral* teleology in the *Groundwork of the Metaphysics of Morals*, the *Critique of Practical Reason*, *Religion within the Boundaries of Mere Reason*, and the *Metaphysics of Morals*.

***

As I have noted already, Kant was the first post-Newtonian philosopher of free will to face up explicitly and fully to the two basic free will problems (namely, free will vs determinism, and Incompatibilism vs. Compatibilism). And I think that everyone who works on either Kant's metaphysics or the free will problem more generally, would agree with that claim. But from there on in, it is not so simple. For it is well known to contemporary Kantians, and especially to contemporary Kantian ethicists, that in scholarly space there exist at least two sharply distinct, competing versions of Kant's theory of freedom, each of which has a fairly solid grounding in Kant's texts: *The Timeless Agency Theory*,[186] and *The Regulative Idea Theory*.[187]

The Timeless Agency Theory adopts the classical Two World or Two Object Theory of the noumena vs. phenomena distinction and asserts that a noumenal subject is autonomous in that it has absolutely spontaneous causal efficacy, or nomological sufficiency, with respect to its self-legislating positively noumenal will, apart from all alien causes and all sensible impulses, from or for the sake of the Categorical Imperative. More precisely, our autonomy consists in our causing, from outside of time and space, phenomenal human behavioral movements (in outer sense) and psychological processes (in inner sense) that are themselves independently necessarily causally mechanically determined by general causal laws of nature plus the settled empirical facts about the past. The Timeless Agency Theory is supported primarily by texts drawn from the *Critique of Pure Reason* (esp. *CPR* A538-558/B566-586).

By contrast, The Regulative Idea Theory adopts the neoclassical Two Aspect or Two Standpoint Theory of the noumena vs. phenomena distinction and says that we are required by our innate capacity for practical reason *to believe or take ourselves to be acting morally only under the rational Idea*—that is, the consistently thinkable, but not empirically cognizable, noumenal representation—*of our own practical freedom or autonomy*. The Regulative Idea Theory is supported primarily by section III of *Groundwork of the Metaphysics of Morals*.

Both The Timeless Agency Theory and The Regulative Idea Theory have some serious problems, however.

On the one hand, it is crucial to note that the texts that best support The Timeless Agency Theory are explicitly said by Kant to demonstrate only the bare conceivability and logical consistency of the notions of freedom and Universal Natural Determinism, and neither the *reality* or actual existence, nor the *real*, non-logical, "strong metaphysical," or synthetic possibility of freedom:

> Do freedom and natural necessity in one and the same action contradict each another? And this we have answered sufficiently when we showed that since in freedom a relation is possible to conditions of a kind entirely different from those in natural necessity, the law of the latter does not affect the former; hence each is independent of the other, and can take place without being disturbed by the other…. It should be noted here that we have not been trying to establish the **reality** of freedom,

> as a faculty that contains the causes of appearance in our world of sense…. Further, we have not even tried to prove the **possibility** of freedom; for this would have not succeeded either, because from mere concepts *a priori* we cannot cognize anything about the possibility of any real ground or any causality. (*CPR* A557-558/B585-586, underlining added, boldfacing in the original)

In turn, the most serious problem with The Timeless Agency Theory is that timeless agency is in fact really, non-logically, strongly metaphysically, or synthetically a priori *im*possible, even though it remains consistently thinkable or conceivable, and thus logically *possible*. If all phenomenal events are independently necessarily determined by natural laws together with antecedent facts, then the noumenal causality of the will implies what is nowadays called the *non-standard causal overdetermination* of phenomenal human behavioral movements in outer sense and psychological processes in inner sense.

The thesis of "non-standard" causal overdetermination says

> (i) there can be two ontologically distinct nomologically sufficient causes of the same event, such that one of the two causes is physical and one of the two causes is *extra-physical*, and each of which can operate in the absence of the other, and correspondingly,
> (ii) there can be two complete and independent causal explanations of the same event.

By contrast, "standard" causal overdetermination postulates two distinct sufficient *physical* causes of the same event, and correspondingly two complete and independent physical causal explanations of the same event. The usual lurid example of standard or physical causal overdetermination is two snipers and one fatal wound, but a more commonplace example is a cautious man's single pair of trousers being completely and independently held up by a belt *and* suspenders. Obviously, this sort of causal phenomenon is not especially philosophically worrying.

In the case of non-standard or extra-physical causal overdetermination, however, as I noted in a slightly different context in section 2.3, and as Jaegwon Kim has compellingly argued, there is a knock-down worry. If there already is a nomologically sufficient physical cause of some event, and if, correspondingly, there already is a complete and independent physical causal explanation of that same event, then this cause and that causal explanation together necessarily *exclude* there being any other distinct, and specifically extra-physical, nomologically sufficient cause or causal explanation of the same event.[188] So the non-standard causal overdetermination implied by The Timeless Agency Theory, although consistently thinkable or conceivable and logically possible, is non-logically, strongly metaphysically, or synthetically a priori necessarily ruled out.

On the other hand, it is also crucial to note that the texts which best support The Regulative Idea Theory are explicitly said by Kant to demonstrate only that "freedom must be presupposed (*vorausgesetzt*) as a property of the will of all rational beings" (*GMM* 4: 447, underlining added) and also only that "all human beings think of themselves as having free will" (*GMM* 4: 455, underlining added), as opposed to our having "cognition" (*Erkenntnis*) of free will or "scientific knowledge" (*Wissen*) of it, both of which are denied to us. Correspondingly, the most serious problem with The Regulative Idea Theory is that even if it were true, it just does

not do the philosophical work required of the noumenal causation *vs*. phenomenal causation distinction. This is precisely because it does not entail either the reality or the real possibility of freedom of the will, but instead entails only at best our *belief* or *faith (Glaube)* in its reality or real possibility. Such an entailment is not only ontologically deflationary but also, arguably, it does not even *epistemically* justify that belief. For according to Kant, as per the Postulates of Practical Reason in the second *Critique*, our belief in freedom is only a self-evident practical belief, not a theoretical belief—in effect, a mode of *rational faith* or "moral certainty" (*CPR* A829/B857). In turn, a belief is held with moral certainty, according to Kant, if and only if it is a sufficiently warranted practical commitment that is nevertheless held by us on "theoretically insufficient" grounds, and "if I know with certainty that no one else can know of any other conditions that lead to the proposed end":

> Only in a **practical relation**, however, can taking something that is theoretically insufficient to be true be called believing (*Glauben*). This practical aim is either that of **skill** or **morality**, the former for arbitrary and contingent ends, the latter, however, for absolutely necessary ends. Once an end is proposed, then the conditions for attaining it are hypothetically necessary. This necessity is subjectively but still only comparatively sufficient if I do not know of any other conditions at all under which the end could be attained; but it is sufficient absolutely and for everyone if I know with certainty that no onel else can know of any other conditions that lead to the proposed end. (*CPR* A823/B851, underlining added, boldfacing in the original)

In other words, this morally certain belief could still be theoretically or scientifically *false*—either in the sense of natural science or physics, or in the sense of the non-natural science of metaphysics—and thus also epistemically wrong, since knowledge is sufficiently justified *true* belief. For all we know, and for all that The Regulative Idea Theory says, then, we could still be nothing but deterministic natural automata and real-world Turing machines, moist robots, or "meat puppets," in the tragic situation of epiphenomenally dreaming that we are free, but with no more causal power of our own than a turnspit.

For these reasons, it seems to me that both The Timeless Agency Theory and The Regulative Idea Theory are very likely to be objectively false, whatever else we may think about the question of which theory most accurately reflects Kant's own considered views about freedom of the will.

In contrast to these theories, as I have mentioned, I want to develop and defend Kant's Biological Theory of Freedom.[189] Just like The Timeless Agency Theory and The Regulative Idea Theory, The Biological Theory also has a fairly solid grounding in Kant's texts, although it is primarily supported by texts drawn from what I have called Kant's "post-Critical" period after 1787,[190] especially including the *Critique of the Power of Judgment* and the *Opus postumum*. But The Biological Theory differs sharply from the other two theories in that

> (i) it avoids their basic serious problems, and
> (ii) arguably it is fairly close to being objectively true.

So I think that we should prefer it both on grounds of inference to the best (namely, most rationally charitable, by our own philosophical lights) interpretation, and also for independent philosophical reasons.

In a nutshell, The Biological Theory says that transcendental freedom of the will in the occurrent sense—that is, deep freedom or up-to-me-ness, as actually occurring—is a far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic process. And the most direct way of recognizing Kant's commitment to The Biological Theory is to see that for him, human persons are rational human *animals*, whose capacity for free will is fully metaphysically continuous with their animality:

> The human being, as animal, belongs to the world, but, as person, also to the beings who are capable of rights—and, consequently, have *freedom* of the will. Which ability essentially differentiates [the human being] from all other beings; *mens* is innate to [the human being]. (*OP* 21:36, underlining added)

This also makes Kant a defender of *liberal naturalism*,[191] which, as we have seen in chapters 1-2, says that the irreducible but also non-dualistic mental properties of rational minded animals are as basic in nature as biological properties, and metaphysically continuous with them. Another name for liberal naturalism is "objective idealism." Objective idealism is sharply distinct both from *subjective idealism*, which says that the world is nothing a phenomenal mental construction of an individual cognizer (as defended, for example, by Berkeley, the neo-Kantians, early Carnap, C.I. Lewis, and Nelson Goodman) and also from *absolute idealism*, which says that the world is nothing but a giant mind, its thought-forms, and its thought-processes (as defended, for example, by Fichte, Schelling, and Hegel). As opposed to either subjective idealism or absolute idealism, liberal naturalism, aka objective idealism, is necessarily equivalent with a commitment to "empirical realism" in Kant's sense and also to an objective, metaphysically modest version of Kant's transcendental idealism that I call *weak or counterfactual* transcendental idealism.[192] The thesis shared by the weak-or-counterfactual-transcendental-idealist Kant, me, and other liberal naturalists is that irreducible, aprioristic, non-instrumental, conscious, intentional, caring, rational, deeply free, agential mindedness *grows naturally* in the manifestly real physical world, in organisms whose lives have an appropriately high level of non-mechanical thermodynamic complexity and self-organization. The manifestly real natural physical world necessarily includes our real possibility and is immanently structured for the dynamic emergence of lives like ours and minds like ours. Or in Nagel's formulation: "rational intelligibility is at the root of the natural order."[193]

In my opinion, Nagel unfortunately overstates his case in this context by immediately comparing his view to those of Schelling, Hegel, and absolute idealism more generally, in the second half of the same sentence:

> The view that rational intelligibility is at the root of the natural order makes me, in a broad sense, an idealist—not a subjective idealist, since it doesn't amount to the claim that all reality is ultimately appearance—but an objective idealist in the tradition … [194] of certain post-Kantians, such as Schelling and Hegel, who are usually called absolute idealists.

But then in the very next paragraph, Nagel re-states his view so that it comes out instead, more carefully and modestly, as, in effect, liberal naturalism and weak or counterfactual transcendental idealism: "nature is such as to give rise to conscious beings with minds; and it is such as to be comprehensible to such beings."

Right, bang on. Nature is *not* nothing but a giant mind. Instead, manifestly real physical nature is such that, necessarily, it is really possible for creatures like us to emerge dynamically, and also such that, were we to exist, then we would be able to perceive it and know it directly, to some non-trivial extent. In effect, liberal naturalism is simply a robust modal inference,

> from (i) the true fact that rational minded animals like us actually exist in the natural world,
> to (ii) our possibility in that natural world (actuality entails possibility),
> to (iii) the necessity of that possibility (possibly P entails necessarily possibly P, as captured and codified in the modal system S5),
> hence: (iv) necessarily, the actual natural world is such that rational minded animals like us are necessarily possible.[195]

That is the view that the weak-or-counterfactual-transcendental-idealist Kant, I, other liberal naturalists, and Nagel all share.

Kant's theory of transcendental freedom is also fundamentally based on his notion of "spontaneity." For him, *X* is spontaneous if and only if *X* is a conscious mental event that expresses some acts or operations of a creature, and *X* is

> (i) causal-dynamically *unprecedented*, in the two-part sense that
> (ia) conscious mental events of those specific sorts have never actually happened before, and (ib) the settled empirical facts about the past together with the general causal laws of nature do not provide nomologically sufficient conditions for the existence or specific character of those conscious mental events,
> (ii) *underdetermined* by external sensory informational inputs, and also by prior desires, even though it may have been triggered by those very inputs or motivated by those very desires,
> (iii) *creative* in the sense of being recursively constructive, or able to generate infinitely complex outputs from finite resources, and also
> (iv) *self-guiding*. (*CPR* A51/B75, B130, B132, B152, A445-447/B473-475)

Furthermore, spontaneity can be either *relative* or *absolute*. Relative spontaneity requires inputs to the rational conscious mind, whereas absolute spontaneity allows the rational conscious mind to generate its own outputs without any triggering inputs. For example, rational human a priori cognition is only relatively spontaneous, because it requires sensory inputs via empirical intuition, whereas an "intellectual intuition," if it existed, would be absolutely spontaneous, because it could cause the objects of its thoughts to exist just by thinking them (*CPR* A19-22/B33-36, B71-72).

Now according to Kant, the concept of a cause analytically entails the concept of its effect, and the general schematized pure concept of CAUSE says that something *X* (the cause) necessitates something else *Y* (its effect) in time according to a necessary rule or law. Or equivalently, according to Kant, to say that *X* causes its effect *Y* is to say that *X* is nomologically sufficient for *Y* in time (*CPR* B112, A144/B183). Then *X* is a relatively or absolutely spontaneous cause of its effect *Y* if and only if

> (i) *X* is nomologically sufficient for *Y* in time, and

> (ii) *X* is a conscious mental event that is necessarily unprecedented, underdetermined by external sensory inputs and desires, creative, and self-guiding.

Finally, relatively or absolutely spontaneous *mental causation* is the same as transcendental freedom:

> By freedom in the cosmological sense … I understand the faculty of beginning a state **from itself** (*von selbst*), the causality of which does not in turn stand under another cause determining it in time in accordance with the law of nature. Freedom in this signification is a pure transcendental idea, which, first, contains nothing borrowed from experience, and second, the object of which cannot be given determinately in any experience…. But since in such a way no absolute totality of [natural] conditions in causal relations is forthcoming, reason creates the idea of a spontaneity, which could start to act from itself, without needing to be preceded by any other cause that in turn determines it to action according to the law of causal connection. (*CPR* A533/B561, underlining added, boldfacing in the original)

Although transcendental freedom is a particularly robust kind of mental causation, in the second *Critique* Kant sharply distinguishes transcendental freedom from mere *psychological* freedom:

> These determining representations [that is, instincts or motives] themselves have the ground of their existence in time and indeed in the *antecedent state*, and in a preceding state, and so forth, these determinations may be internal and they may have psychological instead of mechanical causality, this is, produce actions by means of representations and not by bodily movements; they are always *determining grounds* of the causality of a being insofar as its existence is determinable in time and therefore under conditions of past time, which are thus, when the subject is to act, *no longer within his control* and which may therefore bring with them psychological freedom (if one wants to use this term for a merely internal chain of representations in the soul) but nevertheless natural necessity, leaving no room for *transcendental freedom* which must be thought of as independence from everything empirical and so from nature generally, whether regarded as an object of inner sense in time only or also as an object of outer sense in both space and time; without this freedom (in the latter and proper sense), which alone is practical a priori, no moral law is possible and no imputation in accordance with it. (*CPrR* 5: 96-97, underlining added)

Psychological freedom has two dimensions.

First, in its negative dimension, psychological freedom is the subject's consciousness of choosing or acting without being prevented, and without inner or outer compulsion.

And second, in its positive dimension, psychological freedom is the subject's consciousness of creativity and vitality.

So transcendental freedom is the *spontaneity of consciousness*,[196] whereas psychological freedom is *the consciousness of spontaneity*. As Kant explicitly points out, and as Hume and Leibniz also noted in anticipation of contemporary Compatibilism (aka "Soft Determinism"), it is both logically and metaphysically possible to be psychologically free without being

transcendentally free. This is what Kant very aptly and famously calls "the freedom of a turnspit" (*CPrR* 5: 97), or as I characterized it above, the epiphenomenal dreams of a mere wind-up toy. So psychological freedom is not a *sufficient* condition of transcendental freedom. Nevertheless, according to Kant, psychological freedom remains a *necessary* condition of transcendental freedom, just insofar as it is the veridical, essentially non-conceptual "feeling of life" (see sections 2.3 and 2.4 above)—that is, the pre-reflective first-order consciousness, or subjective experience, of creative, vital animal embodiment, in inner and/or outer sense. No one could be transcendentally free and also at the same time undergo the subjective experience of constantly being prevented from choosing or acting, or of constantly being inwardly or outwardly compelled to choose or act, and of being uncreative and lifeless in those very choices and acts, hence robotic and zombie-like, as, for example, in certain forms of schizophrenia. Indeed, as the Second Analogy of Experience explicitly shows, psychological freedom in both its negative and positive dimensions is necessarily built into the correct mental representation of *any* objective causal sequence, via what Kant calls "the **subjective sequence** of apprehension" (boldfacing in the original), whose ordering is not only always subjectively experienced as "entirely arbitrary" (*ganz beliebig*) and not necessitated, but also as being a direct function of the lively—hence creative and vital—shifts of conscious attention in the rational minded animal:

> In the … example of a house my perceptions could have begun at its rooftop an ended at the ground, but also could have begun below and ended above; likewise I could have apprehended the manifold of empirical intuition from the right or from the left…. [The subjective sequence of apprehension] is entirely arbitrary. (*CPR* A193/B238)

When we correctly ascribe transcendental freedom specifically to the will of a rational human animal or human person, then,

> in addition to (i) the metaphysically positive factor of relative or absolute spontaneity, which confers deep freedom or up-to-me-ness on the person's choices and acts,
> and also in addition to (ii) the phenomenology of psychological freedom, which provides (iia) the subjective experience of being unprevented and uncompelled in one's choices and acts, and also (iib) the subjective experience of being creative and vital in the very same choices and acts, there is also
> (iii) a *metaphysically negative* dimension of freedom which guarantees that the person's choices and acts occur independently of all "alien causes," that is, independently of all pathological inner and unowned outer sources of nomologically sufficient compulsion:

> The will is a kind of causality that living beings have so far as they are rational. *Freedom* would then be that property whereby this causality can be active, independently of alien causes *determining* it; just as *natural necessity* is a property characterizing the causality of all non-rational beings—the property of being determined to activity by the influence of alien causes. The above definition of freedom is *negative*. (*GMM* 4: 446, underlining added)

This is where "practical freedom" comes onto the scene. Practical freedom presupposes but also exceeds transcendental freedom, in that practical freedom is the relative or absolute spontaneity of the will not only independently of all alien causes, but also independently of *all sensible impulses* (that is, empirical, first-order desires):

> **Freedom in the practical sense** is the independence of the power of choice (*Willkür*) from **necessitation** by impulses of sensibility. For a power of choice is sensible insofar as it is pathologically affected (through moving-causes of sensibility); it is called an animal power of choice (*arbitrium brutum*) if it can be **pathologically necessitated.** The human power of choice is indeed an *arbitrium sensitivum*, yet not *brutum*, but *liberum*, because sensibility does not render its action necessary, but in the human being there is a faculty of determining oneself from oneself, independently of necessitation by sensible impulses. (*CPR* A534/B562, underlining added, boldfacing in the original)

As I mentioned above, however, this appeal to practical freedom's necessary underdetermination by empirical, first-order affects is merely a negative characterization of it. As positively characterized, practical freedom also involves the capacity for *self-legislation* in conformity with, and for the sake of, the Categorical Imperative or moral law. Or in other words, practical freedom in the normatively low-bar, qualifying, or capacity sense is also necessarily equivalent with *autonomy* (*GMM* 4: 440-441, 446-463) in the normatively low-bar, qualifying, or capacity sense.

It may seem, on the face of it, that there is and should be no direct connection whatsoever between the person's relatively or absolutely spontaneous, psychologically free, autonomous will, on the one hand, and her actual existence in physical nature on the other. That is the basic idea behind the ontologically dualist, Classical agent-causal Libertarian theory, according to which the freely willing person, as an agent-cause, necessarily stands outside the natural causal order of events in spacetime.[197] Indeed, Kant is often cited as a paradigmatic defender of the Classical agent-causal Libertarian theory—as per The Timeless Agency Theory.[198] But in fact Kant himself explicitly asserts otherwise:

> Practical freedom can be proved through experience. For it is not merely that which stimulates the senses, i.e., immediate affects them, that determines human choice, but we always have a capacity to overcome impressions on our sensory faculty of desire by representations of that which is useful or injurious even in a more remote way; but these considerations about that which in regard to our whole condition is desirable, i.e., good and useful, depend on reason. Hence this also yields laws that are imperatives, i.e., objective **laws of freedom**, and that say **what ought to happen**, even though it never does happen…. We thus cognize practical freedom through experience, as one of the natural causes, namely a causality of reason in the determination of the will. (*CPR* A802-803/B830-831, underlining added, boldfacing in the original)
>
> Two things fill the mind with ever new and increasing admiration and reverence, the more often and more steadily one reflects on them: *the starry heavens above me* [that is, nature] *and the moral law within me* [that is, freedom]. I do not need to search for them and merely conjecture them as though they were veiled in obscurity or in the

> transcendent region beyond my horizon; I see them before me and connect them immediately with the consciousness of my existence. (*CPrR* 5: 161-162)
>
> Now although there is an incalculable gulf fixed between the domain of the concept of nature, as the sensible, and the domain of the concept of freedom, as the supersensible…: yet the latter **should** have an influence on the former, namely the concept of freedom should make the end that is imposed by its laws real in the sensible world; and nature must consequently also be able to be conceived in such a way that the lawfulness of its form is at least in agreement with the possibility of the ends that are to be realized in it in accordance with the laws of freedom. (*CPJ* 5: 176, underlining added, boldfacing in the original)

In other words, Kant is explicitly saying that transcendental freedom is not only really *possible* but also even real, *full-stop*, that is, actually and really, "as one of the natural causes" that is sensibly and "immediately," hence intuitionally and essentially non-conceptually, verified, in this case by "the consciousness of my existence." So for Kant there is a *natural piety*[199] *Cogito*:

> I feel dual reverence for outer nature and inner morality, therefore I exist as a free, real cause of natural events.

I will now reconstruct Kant's reasoning for this perhaps very surprising thesis, and in so doing, argue that his theory of transcendental freedom can be very plausibly—and philosophically most defensibly—interpreted as a biologically-based, Kantian Non-Conceptualist theory. As I mentioned above, I shall be drawing primarily on texts from Kant's post-Critical period after 1787, and in particular from the third *Critique*.

As we saw in sections 2.3 and 2.4 above, Kant argues in the two Introductions to the *Critique of the Power of Judgment* and again in the second half of the book, that the concepts LIFE and ORGANISM, and in particular the concept of a "natural purpose" (*Naturzweck*) or living organism, are not ordinary empirical concepts applying to naturally-mechanized physical matter, or to naturally-mechanized material objects, and that they invoke a type of causation that cannot be cognized or known by means of any natural science or *Naturwissenschaft*, where the paradigm of a natural science or *Naturwissenschaft* is classical Newtonian deterministic-mechanistic mathematical physics:

> For a body to be judged as a natural purpose in itself and in accordance with its internal possibility, it is required that its parts reciprocally produce each other, as far as both their form and their combination is concerned, and thus produce a whole out of their own causality, the concept of which, conversely is in turn the cause (in a being that would possess the causality according to concepts appropriate for such a product) of it in accordance with a principle; consequently the connection of **efficient causes** could at the same time be judged as an **effect though final causes**. In such a product of nature each part is conceived as if it exists only **through** all the others, thus as if existing **for the sake of the others** and **on account** of the whole, i.e., as an instrument (organ), which is, however, not sufficient (for it could also be an instrument of art, and thus represented as possible at all only as a purpose); rather it must be thought of as an organ that **produces** the other parts (consequently each produces the others reciprocally), which cannot be the case in any instrument of art, but only of nature,

> which provides all the matter for instruments (even those of art): only then and on that account can such a product, as an **organized** and **self-organizing** being, be called a **natural purpose**. (*CPJ* 5: 373-374, underlining added, boldfacing in the original)

> Strictly speaking, the organization of nature is … not analogous with any causality that we are cognitively acquainted with (*kennen*). (*CPJ* 5: 375, underlining added)

Now because the causality of living organisms is uncognizable and unknowable by means of deterministic-mechanistic natural science, the basic concepts of biology are merely "regulative" or "hypothetical" concepts of *reason*, that is, heuristic, logical-fictional, or analogical concepts employed in our everyday encounters with the natural world, for purposes of the unification and promotion of natural scientific inquiry (*CPJ* 5: 369-415; see also *CPR* A642-647/B670-675).[200] This makes the notion of the causality of living organisms at best a merely *conceptual*, epistemological, and projectivist notion, and not a metaphysical notion. For example, Korsgaard writes:

> [T]he way we conceptualize the world, the way we organize it into a world of various objects, guarantees that it will appear to be teleologically organized at the level of those objects.
>
> Nothing I'm saying here is incompatible with a Darwinian account of how the world became populated with items fit to be thus conceptualized.[201]

In other words, the regulative use of the concept of the causality of living organisms implies a Two-Aspects-Theory-style Compatibilism, that is, a Kantian version of Soft Determinism, via asserting, on the one hand, the pragmatic usefulness or even the practical necessity of *teleological judgments based on teleological concepts*, and on the other hand, the truth of Universal Natural Determinism and Natural Mechanism.

But here is another place where I significantly part company with standard contemporary interpretations of Kant's Critical philosophy in general and of his philosophy of biology in particular. Suppose it is true that Kant is committed to the thesis that teleological judgments and teleological concepts are only regulative, not constitutive. Nevertheless, on fully Kantian (Non-Conceptualist) grounds we can still assert that organismic life—and in particular, the organismic life of my own animal body—is directly cognized by essentially non-conceptual, non-propositional, or non-epistemic, and thus non-judgment-based means. Indeed, as I mentioned in section 2.5 above, according to Kant in the First Part of the third *Critique*, the feelings of pleasure and pain, bodily affects including bodily desires and drives, and proprioceptive feelings, jointly constitute "the feeling of life" (*CPJ* 5: 204, 278), or the feeling of embodied vitality. Now the feeling of life is *primed* by the judgment of taste, but not *constituted* by the judgment of taste, Furthermore, as we also saw in sections 2.3, 2.4, and 2.5 above, according to Kant there is an essential connection between the affective-emotional psychological life of my mind, and the biological life of my own body:

> [L]ife is the subjective condition of all our possible experience. (*Prol* 4: 335)

> Life without the feeling of the corporeal organ is merely consciousness of one's existence, but not a feeling of well- or ill-being, i.e., the promotion or inhibition of the

> powers of life; because the mind for itself is entirely life (the principle of life itself), and hindrances and promotions must be sought outside it, though in the human being himself, hence in combination with his body. (*CPJ* 5: 278, underlining added)

So in these ways, according to Kant, biological life is metaphysically continuous with the essentially embodied conscious, intentional, caring, rational human mind. Otherwise put, our essentially non-conceptual, phenomenal, affective-emotional consciousness in inner sense necessarily entails the existence of biological life as far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, minded animal, thermodynamic process. So for Kant there is also an *animal faith*[202] *Cogito*:

> I feel things "in my bones," therefore I exist as a minded living organism.

In other words, conscious, intentional, caring, rational human beings are, necessarily, *also* living organisms.

This is a crucially important point that bears repeating. The semantic and epistemic constraints that Kant explicitly places on teleological *judgments* and *concepts* about distal material objects in space in the context of biological science as he understood it—namely, that such judgments involving such concepts are always regulative and not constitutive—do *not* apply to the human conscious experience of living organismic embodiment. This subjective experience is essentially non-conceptual or intuitional, and affective-emotional in character, and therefore essentially not conceptual, propositional, epistemic, or judgment-based. So, again, I think that there is a fundamentally important Kantian distinction to be drawn between

> (i) teleological *judgment*, which is inherently neither directly referential nor existentially-committed, because it is inherently based on concepts and also regulative, and
> (ii) teleological *consciousness* or *subjective experience*, in the special mode of essentially non-conceptual contents or *intuitions*, in inner sense, which are inherently both directly referential, and also existentially committed, hence veridical.

According to my reading of Kant's view, more precisely then, I actually have veridical, teleological, inner and outer sense intuitions (that is, veridical, teleological, essentially non-conceptual conscious or subjective experiences) of my own organismic, value-laden, goal-directed, embodied biological life as a (rational) human minded animal. In this way, even if teleological judgments and concepts have only a regulative use, and even if the truth of judgments about natural purposes and the truth of mechanistic, deterministic Newtonian physical science as a theory are incompatible—which is the source of The Dialectic of Teleological Judgment—nevertheless, I still can, and in fact still really do, have an essentially non-conceptual, non-propositional, non-epistemic, non-judgment-based teleological *biophenomenology* that is veridical and therefore constitutive.

It follows from all this that there are real organismic, value-laden, goal-directed biological facts in physical nature, amongst which are all the mental facts of minded animal life. It is just that I cannot *scientifically know* these biological and biophenomenological facts via physical theories that are exclusively based on naturally mechanistic, deterministic principles. But I can

still *truly consciously feel* at least some of these biological and biophenomenological facts, by truly consciously feeling, "in my bones," my own essentially embodied minded animal life. Furthermore, and most importantly of all for the present purposes of argument, by way of these teleological intuitions or veridical teleological essentially non-conceptual cognitions, according to Kant, and according to me, I can also truly consciously feel my own transcendental freedom:

> *Sensible life* has, with respect to the *intelligible* consciousness of its existence, (consciousness of freedom), the absolute unity of a phenomenon, which, so far as it contains merely appearances of the disposition that the moral law is concerned with (appearances of the character), must be appraised not in accordance with the natural necessity that belongs to it as appearance but in accordance with the absolute spontaneity of freedom. (*CPrR* 5: 99, underlining added)

This is natural piety and animal faith *in action*. In other words, Kant's natural piety Cogito and his animal faith Cogito turn out to be one and the same.

In view of these points, as I am understanding him, Kant regards empirical psychology as a constitutive and nomological yet also *non*-deterministic and *non*-mechanistic "life science" of the human minded animal. Otherwise put, empirical anthropology and empirical psychology jointly constitute one Kantian "moral science" (*Geisteswissenschaft*)—call it "empirical anthropsychology"—that is inherently *not* a mechanistic, deterministic Newtonian mathematical physics, or "natural science" (*Naturwissenschaft*). To be sure, Kantian empirical anthropsychology contains unique "psycho-psycho" laws that strictly govern the real phenomenological facts of inner sense,[203] which, we now recognize, must also be real *bio*phenomenological facts. Nevertheless, these biophenomenological facts in empirical anthropsychology cannot be arithmetically analyzed:

> The empirical doctrine of the soul must always remain … removed … from the rank of what may be called a natural science proper. This is because mathematics is inapplicable to the phenomena of the inner sense and their laws…. It can, therefore, never become anything more than a historical (and, as such, as much as possible) systematic natural doctrine of the inner sense, i.e., a natural description of the soul, but not a science of the soul. (*MFNS* 4: 471, underlining added)

Moreover, I believe that we can also charitably reconstruct Kant's rationale for holding this important thesis, in the following way.

As we have already seen, the merely subjective temporal ordering of conscious states, or subjective experiences, in inner sense is "entirely arbitrary" (*ganz beliebig*) (*CPR* A193/B238), according to the desires and choices of the conscious, intentional, caring, rational human animal. The radical open-endedness of possible orderings in inner sense, in turn, means that the set of all mental biophenomena cannot be put into a one-to-one correspondence with the set of natural numbers, or reconstructed as Turing-computable functions of Primitive Recursive Arithmetic. Kant's interestingly restricted (pre-Peano, pre-Frege) conception of arithmetic,[204] together with the Axioms of Intuition and the Anticipations of Perception—that is, the mathematical synthetic a priori principles of pure understanding (*CPR* A160-162/B199-201)—and the Analogies of Experience, show that the mechanistic theory of Universal Natural Determinism, as Kant understood it, requires the simple primitive recursive arithmetization of

causal processes in time. But as we saw in section 2.4 above, given Gödel's incompleteness theorems together with The Church-Turing Thesis, it follows that mathematical truth, whether in Peano Arithmetic or dynamical systems theory, is inherently uncomputable or non-mechanical. Furthermore, given my Leibniz-inspired and Searle-inspired Arm-Waving Room Argument in section 2.2 above, it follows that the intentional body-movements of rational human animals instantiate inherently uncomputable and non-mechanical biological and biophenomenological functions. Thus for Kant, and also correspondingly for me, a neo-Aristotelian and contemporary Kantian dynamicist, the anthropsychological laws of rational human animal life cannot be either deterministic or more generally naturally mechanized. Furthermore since essentially embodied human mental life entails biological life, it follows directly from Kant's mind-in-life thesis and contemporary dynamical systems theory, together with Gödel's incompleteness theorems and The Church-Turing Thesis, together with The Arm-Waving Room Argument, that that not only can there never be a Newton of a blade of grass, but also even more importantly that *there can never be a Newton, Church, or a Turing* of either biological life or the rational human minded animal.

So yet again, we can clearly see that our rational human minded animal life, especially including the exercise of our power of choice or *Willkür*, cannot be naturally mechanized,[205] precisely because it is an inherently uncomputable, self-organizing, living organismic process, and precisely because its immanent structure, our conscious freedom, is inherently a form of life.

There is an instructive contrast here to be drawn between what I am claiming on the one hand, and J.R. Lucas's famously controversial thesis that Gödel's incompleteness results entail the existence of freedom of the will on the other.[206] It is to be particularly emphasized that I do not accept Lucas's argument that Gödel's incompleteness results, on their own, without supplementation by several other substantive epsistemic and metaphysical premises, entail the existence of free will: that is too strong a thesis. All I am accepting is Lucas's subsidiary thesis that, from Gödel's incompleteness results, *plus* our authentic a priori knowledge of some strictly logically unprovable (in any *Principia Mathematica*-style logical system, plus enough axioms of Peano Arithmetic), but still rationally intuitable truths of mathematics—for example, arguably,[207] Goldbach's conjecture, and the Continuum Hypothesis—it follows that the rational human mind is not essentially a digital computer. And from that thesis, together with the mind-in-life thesis, the non-deterministic interpretation of non-equilibrium/complex systems thermodynamics, the dynamicist model of life, and the concept of Natural Mechanism, it also follows there exist some non-mechanical natural processes, namely precisely those natural mental processes that are constitutive of our authentic a priori knowledge of some logically unprovable truths of mathematics.

How does this all apply to Kant's theory of transcendental freedom? The answer is, that according to The Biological Theory of Freedom, even if all the inert, non-living parts of macroscopic material nature, as metaphysically described by the three Analogies of Experience, do indeed fall under the general causal laws of classical mechanistic, deterministic Newtonian mathematical physics, nevertheless the existence of these deterministic natural automata is fully consistent with and indeed is also metaphysically presupposed by the instantiation of an irreducibly different set of properties in the special (but not in any way supernatural) kind of living organism that is the rational human animal. This is a set of irreducible, non-dualist, non-supervenient, inherently non-mechanical, uncomputable, conscious, intentional, caring, a priori, and categorically normative properties, whose precise pattern of instantiations

constitutes both that animal's power of choice and also its transcendental and practical freedom of the will, or its autonomy. And it also brings into existence ontologically and dynamically emergent, complex, self-organizing, living organismic thermodynamic processes, and also conscious, intentional, caring, spontaneously causally efficacious, nomologically one-off, or one-time-only, facts about the essentially embodied agency of rational human minded animals.

These natural facts about rational human animal body movements are thereby *non-locally compatibilistic* but also *locally incompatibilistic* in the ways I spelled out in section 1.2 above. That is, none of the general causal laws of nature are ever violated by these animal movements, and in fact they are actually presupposed by these animal movements, but at the same time, neither the existence nor the specific character of these animal movements in the present or the future is entailed or necessitated by all the general causal mechanical laws, especially including the Conservation Laws, together with the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang. And that is precisely because these rational human animal body movements are caused by our transcendental freedom, which is a non-empirical or a priori but still fully natural biological and neurobiological fact about us. In this way, rational human animals are *not* deterministic (or for that matter, indeterministic) natural automata or machines, and correspondingly they therefore *are* real persons. Indeed, in the *Critique of Practical Reason* Kant explicitly asserts that rational personhood (*Persönlichkeit*) itself is essentially

> freedom and independence from the mechanism of nature regarded as a capacity of a being subject to special laws (pure practical laws given by its own reason). (*CPrR* 5: 87, underlining added)

In this way, the source-incompatibilistic difference between, on the one hand, (i) the general causal mechanical laws of nature, and on the other hand, (ii) the non-deterministic, non-mechanical, one-off or one-time-only laws of rational human minded animal intentional body movement, guided by categorically normative moral laws, is itself perfectly consistent with the thesis that these two different types of laws are fully *compatible* in a world in which type-(i) laws apply to all and only things that strictly obey the Conservation Laws, Big-Bang-causation, and Turing computable algorithms, and more generally are naturally deterministic or indeterministic processes, whereas type-(ii) laws apply to all and only rational minded animals, or real persons, is itself the metaphysical core of Kant's Biological Theory of Freedom. This perfectly consistent conjunction of Local Incompatibilism together with Non-Local Compatibilism, insofar as it is entailed by Kant's Biological Theory of Freedom, is what I also call "Kant's Incompatibilistic Compatibilism."

3.3 PRACTICAL-FREEDOM-IN-LIFE: STRONG KANTIAN NON-INTELLECTUALISM

Let us now suppose, for the purposes of argument, that Kant's Biological Theory of Freedom and his Incompatibilistic Compatibilism, as I have presented them, are honest-to-goodness *true*, not only as charitable interpretations of Kant's theory of free will but also, far more importantly, objectively speaking, quite apart from textual interpretation. That fixes my

"third way" approach to Kant's, and also to a contemporary Kantian, metaphysics of free will. What I want to do in this section and the next, building directly on the neo-Aristotelian and contemporary Kantian anti-mechanist, non-reductive, dynamicist philosophy of biology I put in place in the last chapter, together with Kant's Biological Theory of Freedom and his Incompatibilistic Compatibilism as I have spelled them out in the just-previous section, is to work out a corresponding contemporary Kantian theory of *free agency*, as practical-freedom-in-life.

It is also directly relevant to note in this connection that I have already presented a full-scale action-theory elsewhere,[208] including concrete examples, critical analyses of the relevant secondary literature, etc., as well as corresponding theories of mental causation and also of the key role of the emotions in intentional action, that is presupposed by the current discussion. The current discussion is intended primarily to bring out the specifically *contemporary Kantian* dimensions of that full-scale action-theory.

According to the classical and standard reading of Kant's practical philosophy and his theory of practical agency—defended, for example, by Sidgwick and Korsgaard[209]—Kant is an *intellectualist* who believes that the rational human innate capacity for conceptualization and self-conscious thinking, the understanding, necessarily determines all the intentional contents of the will and practical reasoning, whether this is instrumental practical reasoning via the hypothetical imperative, that is, via the *impure* rational will or *Wille*, or non-instrumental practical reasoning via the Categorical Imperative, that is, via the *pure* rational will or *Wille*.

One fundamental thing to recognize here is that all Kantian intellectualists about practical intentional content are also, perforce, defenders of Kantian Conceptualism about cognitive intentional content, which says that the rational human innate capacity for conceptualization and self-conscious propositional thinking, *the understanding* (*Verstand*), necessarily determines all the intentional contents of cognition.[210] More precisely, all Kantian *intellectualists* are also Kantian *conceptualists*, because representational contents of the sorts captured by judgments in the Kantian sense[211]—whether directly referential contents, conceptual contents, propositional contents, or formal-logical contents—are all essentially the same sorts of contents, whether in the context of theoretical judgments and theoretical inferences or in the context of practical judgments and practical inferences. Hence if, as I believe and have also argued here and elsewhere (see, for example, section 2.4 above), Strong Kantian Non-Conceptualism is true, then Kantian intellectualism is false from the get-go.

By sharp contrast, then, what I want to argue in this section is that Kant is a *strong non-intellectualist* who believes that *not all* the intentional contents of practical reasoning and the rational will, whether instrumental or non-instrumental, are necessarily determined by the understanding, and also that *at least some* of the intentional contents of practical reasoning, both instrumental and non-instrumental alike, are necessarily determined by the rational human innate capacity for sensibility in a practical sense, in an essentially non-conceptual, non-self-conscious, non-propositional, and non-inferential way. This capacity for practical sensibility includes the affective sub-capacities for conation, desire, emotion, and feeling, as well as the power of choice, the sensible will, or *Willkür*—in effect, it is nothing more and nothing less than the human heart. So Kant's strong non-intellectualism says that free choice and autonomous willing *can come straight from the human heart*. Or as Kant himself puts it:

> The capacity or incapacity of the power of choice/sensible will (*Willkür*) that arises from this natural propensity to adopt or not to adopt the moral law in its maxims can be called *the good or evil heart*. (*Rel* 6: 29, underlining added)

It is also particularly to be noted that Kant's non-intellectualism includes, but is not restricted to, what is nowadays called the *affectivist* interpretation of Kant's theory of moral motivation.[212] This is because Kant's strong non-intellectualism, as I am formulating and defending it, is not *merely* the thesis that our rational human affective capacities can directly motivate free choice and autonomous moral agency (= affectivism), but instead the *two*-part thesis which says

> (i) that our rational human affective capacities can directly motivate free choice and autonomous willing (= affectivism), and
> (ii) that our human affective capacities can directly motivate free choice and autonomous willing *only insofar as* they necessarily determine the intentional contents of practical reasoning and the rational will in an essentially non-conceptual, non-self-conscious, non-propositional, and non-inferential way (= Strong Kantian Non-Conceptualism as applied to Kant's theory of practical agency).

In this way, Kant's strong non-intellectualism is the conjunction of affectivism and Strong Kantian Non-Conceptualism, as applied to Kant's theory of practical agency. Or in other words, for Kant free choice and autonomous willing can come not only from the human heart, and not from the head, understanding, or intellect, but also, insofar as they *do* come from the human heart, only *straight* from the heart, and not indirectly mediated by concepts, self-conscious deliberation, propositional reasons, or inferences.

Fully explicitly and specifically then, according to what I am calling *Strong Kantian Non-Intellectualism*, all rational human practical and moral action not only begins in our specifically human sensibility (since that can be accommodated by the weaker versions of Kantian Conceptualism and Kantian intellectualism alike, according to which sensibility is a causally necessary but non-constitutive condition of cognitive intentionality or practical intentionality[213]), but also

> (i) it is primitively *grounded on* the first-order essentially non-conceptual, non-self-conscious, non-propositional, and non-inferential affective intentional inputs deriving from empirical conation, desire, emotions, feeling, and choice/sensible willing, and
> (ii) it is necessarily *limited by* the higher-order essentially non-conceptual, non-self-conscious, non-propositional, and non-inferential affective intentional inputs deriving from the non-empirical so-called "Fact of Reason" or *Faktum der Vernuft* (*CPrR* 5: 31; see also 5: 6, 42-43, 47-48, 55-57, 91-94, and 104-108), which is essentially the conscious manifestation of a capacity for *loving* the moral law that is innately *inscribed* in the human heart (*Rel* 6: 84, 145).

According to Strong Kantian Non-Intellectualism, The Fact of Reason is merely "so-called," because it would be far more appropriately called *The Affect of Reason*, insofar as it is most accurately construed as a conscious manifestation of higher-order, essentially non-

conceptual, non-self-conscious, non-propositional, non-inferential conation, desire, emotion, and feeling (aka the affects). More precisely, The Fact/Affect of Reason is fundamentally expressed as the higher-order, essentially non-conceptual, non-self-conscious, non-propositional, non-inferential, and also specifically moral, affects of

> (i) "respect" or *Achtung*, and
> (ii) "self-fulfillment" or *Selbstzufriedenheit*,

both of which in turn are specifically moral conscious manifestations of what I will call *the desire for self-transcendence*. According to my account, the desire for self-transcendence is *the fundamental higher-order desire to be moved to choice and action by non-egoistic, non-hedonistic, non-consequentialist first-order desires*.

I will discuss self-fulfillment or *Selbstzufriedenheit* later, in section 3.4. But for now it is most important to note that the desire for self-transcendence, consciously manifested as the higher-order, essentially non-conceptual, non-self-conscious, and non-propositional affect of respect, that is, a love of the moral law that is innately inscribed in the human heart, "is properly the representation of a worth that infringes upon my self-love (*Eigenliebe*)" (*GMM* 4: 402n.). Correspondingly, the directly-referential object or target of that representation is *the moral law*, namely, *the Categorical Imperative*, innately specified in the hearts of human persons, who inherently possess the absolute intrinsic non-denumerable objective moral value of *dignity*, and are inherently capable of transcendental and practical *freedom*. Therefore, that target

> is at the same time an object of *respect* inasmuch as … it *weakens* self-inflation (*Eigendünkel*); and inasmuch as it even *strikes down* self-inflation, that is, humiliates it, it is an object of the greatest *respect* and so too the ground of a positive feeling that is not of empirical origin and is cognized a priori. (*CPrR* 5: 73)

Morally true love overcomes self-love. In this way, via the higher-order essentially non-conceptual, non-self-conscious, non-propositional, and non-inferential feeling of respect, the desire for self-transcendence directly opposes both "self-love" or *Eigenliebe* (roughly, narcissism) and also "self-inflation" or *Eigendünkel* (roughly, selfishness), which is to say that it directly opposes human *egoism* in all its forms.[214]

An essential feature of the higher-order desire for self-transcendence, as we will see, is that it can robustly manifest itself not only in morally *right* choice and action (as the higher-order, essentially non-conceptual, non-self-conscious, non-propositional, and non-inferential affects of respect and self-fulfillment), but also in morally *wrong* choice and action, even to the point of umitigated evil and wickedness. At this point, paradoxically, the desire for self-transcendence, as the higher-order desire that inherently opposes first-order willing that is egoistic, hedonistic, or consequentialist, can turn into its dialectical opposites, self-love or narcissism and self-inflation or selfishness, by becoming the monstrous narcissism and selfishness of Hume's fictional person who would allow the whole world and everyone in it, including himself, to be destroyed just so that he could avoid having his finger scratched,[215] and also of Milton's fictional fallen angel, Satan, who chillingly says:

> So farewell, hope; and with hope farewell, fear; Farewell, remorse! all good to me is lost; Evil, be thou my good.[216]

Now I do not mean to say that any actual or possible rational human animal could ever be *wholly* Satanic—indeed, Kant explicitly rejects this (*Rel* 6: 35), and I completely agree—but rather only that some rational human choices and acts really are *near*-Satanic, and also that some people have characters that really are near-Satanic. Outside of Miltonic poetic fiction and in the real world, the "human, all too human" capacity for freely choosing the bad and the wrong, whether *as*-bad-and-wrong, or seemingly as-good-and-right, is what Augustine calls "the perversity of the will,"[217] and what Kant, clearly following Augustine, calls "the perversity of the human heart" (*Rel* 6: 30, 37). In view of Kant's full recognition of the perversity of the will/the human heart, then,

> (i) the fact that near-satanic, monstrous narcissism and selfishness are humanly actual and really possible (for example, Hitler and Stalin), together with
> (ii) the all-too-familiar fact of everyday self-love/narcissism and self-inflation/selfishness,

collectively yield the dual fact of what Kant calls *radical evil* (*Rel* 6: 18-33). Radical evil, the perversity of the human will or heart, is the fully natural and "human, all too human" disposition, tendency, or *Gesinnung*, towards freely-chosen *moral disvalue or wrongness*, whose fundamental cause is human egoism in all its forms.

In this connection, it is crucially important to note that in the contemporary Kantian scheme of moral rational normativity that I am developing and defending here, there are two fundamentally different kinds of moral disvalue or wrongness (which is the same as Kant's "radical evil"—hence my use of the term "evil," aka *Böse*, is narrower than Kant's):

> (1) *moral evil*, which is choice or action involving the intentional violation of people's dignity, that is, considering or treating them like things, like mere instruments, or, even worse, like garbage or offal, and
> (2) *non-evil moral badness*, aka *das Übel*, aka *schlecht*, which is the non-evil *privation*, or falling-short-of, ideal or high-bar good, for example, in choices or acts involving benevolence or kindness to others, or sensitivity to their needs, and the related thought that "we can never do enough to help others."

According to this normative scheme, moral evil and moral badness are inherently *lexically ordered* in relation to moral disvalue or wrongness. Clearly, moral disvalue or wrongness, just like moral value or rightness, always comes in degrees: morally wrong choices and acts are always more or less so, just as morally right choices and acts are. Other things being equal, it is much worse literally to stab someone with the intention of murdering her, than it is merely to say "cutting" things to her with the intention of hurting her feelings. But according to this lexical ordering between moral evil and non-evil moral badness, even the least case of real moral evil is fundamentally worse than even the greatest case of real non-evil moral badness. The person who rejoices in the suffering of another, or who acts specifically in order to make somone else suffer, is fundamentally worse than even the biggest con artist, embezzler, or thief that you can think of, although obviously the utilitarian bad consequences of the latter's choices and acts can massively outweigh those of the former.

Again, it is one morally disvalued thing to fall short of the best you can be, for example, as regards benevolence and kindness to others, and sensitivity to their needs (moral badness), but radically another to violate people's dignity (moral evil). Non-evil morally bad choices and acts are all about human imperfection and weakness, that is, being "human, all too human," whereas evil choices or acts strike at the heart of human real personhood itself. It is also very important to note that, corresponding to the overarching normative distinction between moral evil and non-evil moral badness, under the same rubric of moral evil, there are in fact two sharply different further sub-kinds of moral evil:

> (1i) *near-Satanic evil*, that is, evil chosen or done for its own sake, whatever the consequences, as the result of *titanic egoism*—for example, Hitler, and
> (1ii) *banal evil*, that is, evil chosen or done for merely self-interested reasons (aka "banal egoism"), for hedonistic reasons, or for consequentialist reasons—for example, the choices and acts of Adolf Eichmann, aka "the man in the glass booth," as per Hannah Arendt's famous moral analysis of the Eichmann trial in 1961.[218]

Arendt's basic (and, I think, ultimately Kantian) point, with which I also completely agree, is that on the assumption that we hold the stunning general fact of the moral horror of Nazism temporarily fixed for the purposes of some further moral reflection, there is *still* an intrinsic moral difference between Hitler's kind of evil and (Arendt's) Eichmann's kind of evil.

I will come back to all these vitally important distinctions again later in the chapter. But for the moment I want to emphasize that the contemporary Kantian approach to moral rational normativity that I am developing and defending here will doubtless come as a big surprise to many philosophers and non-philosophers alike, who regard Kant as the Puritanical enemy of all human affect, conations, desires, emotions, and feelings. As Ido Geiger aptly notes in a slightly different but still closely-related connection,

> [t]his must come as a great surprise to those readers who take the caricature of Kant as the mortal enemy of all human feelings to be a realistic portrait of his views.[219]

In other words, contrary to a classical and standard interpretation of Kant's ethics and theory of practical agency that locates the ground of all wrong choice and action in affect, conation, desire, emotion, and feeling—that is, in practical *sensibility*—in fact, according to Kant in *Religion within the Boundaries of Mere Reason*, and according to me too, it is *human egoism* that is "the root of all evil," in the broad sense that it is the motivational ground of what I am calling "moral evil" (whether near-Satanic or banal) and of "non-evil moral badness" alike, by virtue of its being the ultimate source of what Kant calls our "self-incurred perversity" of the will. Indeed, for Kant in *Religion*,

> *[c]onsidered in themselves* natural inclinations are *good*, i.e., not reprehensible, and to want to extirpate them would not only be futile but harmful and blameworthy as well (*Rel* 6: 58, underlining added), and

> there is absolutely no salvation for human beings except in the innermost adoption of genuine moral principles in their disposition, [and] … to inferere with this adoption

> is surely the not so often blamed sensibility but a self-incurred perversity or, as we might otherwise also call this wickedness, fraud (*faussité*, the satanic guile through which evil comes into the world): [this is] a corruption that lies in all human beings and cannot be overcome except through the idea of the moral good in its absolute purity. (*Rel* 6: 83)

Part of this bears repeating because it might otherwise seem so shockingly unKantian as to cloud philosophical recognition of its actually being what Kant is asserting:

> *[c]onsidered in themselves* natural inclinations are *good*, i.e., not reprehensible, and to want to extirpate them would not only be futile but harmful and blameworthy as well (*Rel* 6: 58, underlining added).

Considered in themselves, natural inclinations are good! Therefore in his all-things-considered ethical theory circa 1792 (the year of the publication of the first edition of *Religion*), Kant is *not* a philosophical enemy of practical sensibility. On the contrary, he is explicitly a philosophical enemy of those who are philosophical enemies of practical sensibility. According to Kant's considered view, then, he is no more an enemy of practical sensibility than he is an enemy of cognitive sensibility. On the contrary: Kant's considered view in theoretical and practical philosophy is consistently and globally *sensibility first*. This in turn means that not only the standard *intellectualist reading* of Kant's theory of practical agency, but also the standard *non-intellectualist critique* of Kant's theory of practical agency, are equally misguided.[220] Indeed, Kantian ethics, properly understood, and "the ethics of care," properly understood, *are one*.

Correspondingly, and perhaps most surprisingly of all, *pure* practical sensibility, that is, *moral* sensibility—epitomized as The Fact/Affect of Reason, that is, the feeling of respect, that is, the true love for the moral law that is inscribed in the human heart, and self-fulfillment—is the Kantian motivational ground of radical goodness. And that, in turn, brings out the core of truth in Derek Parfit's seemingly very surprising observation, quoted as the last epigraph of this chapter, that Kant is an "emotional extremist." Parfit intends this to be a serious criticism of Kant's theory of practical agency; *contrariwise*, I am taking it to be the essence and greatest strength of Kant's theory.

As we saw above, practical freedom is defined by Kant in a negative way as the independence of first-order volition, or the "power of choice" (*Willkür*), from necessitation by sensible impulses:

> **Freedom in the practical sense** is the independence of the power of choice (*Willkür*) from **necessitation** by impulses of sensibility. For a power of choice is **sensible** insofar as it is pathologically affected (through moving-causes of sensibility); it is called an animal power of choice (*arbitrium brutum*) if it can be **pathologically necessitated**. The human power of choice is indeed an *arbitrium sensitivum*, yet not *brutum*, but *liberum*, because sensibility does not render its action necessary, but in the human being there is a faculty of determining oneself from oneself, independently of necessitation by sensible impulses. (*CPR* A534/B562, boldfacing in the original)

But practical freedom, on its positive side, is also necessarily equivalent to the *realization* of Kantian autonomy:

> The moral law expresses nothing other than the *autonomy* of pure practical reason, that is, [practical] freedom (*CPrR* 5: 33)

More precisely, however, practical freedom, or realized autonomy, is how a transcendentally free person actually chooses or does things by means of her subjective experience or consciousness of recognizing the Categorical Imperative or moral law as a desire-overriding, strictly universal, a priori, categorically normative, non-instrumental practical reason that has both motivating and justifying force. The actual fact of this subjective experience or consciousness of realized autonomous agency is, as I have mentioned, what Kant calls the "Fact of Reason" (*Faktum der Vernunft*) (*CPrR* 5: 31; see also 5: 6, 42-43, 47-48, 55-57, 91-94, and 104-108). Using The Fact of Reason as the key notion, I will presently argue that a Kantian theory of practical freedom, or realized autonomy, entails a special form of *internalism about practical reasons* that shares some non-trivial features with Hume's theory of practical reasoning. But as a preliminary to that, I will say something in general about reasons.

It is plausible to hold that reasons are (or are provided for agents by) facts that *normatively support*—that is, motivate and/or justify (or: legitimate, warrant, etc.)—intentional aims and actions or cognitive beliefs, and do not merely *cause or mechanically trigger* those aims, actions, or beliefs. Reasons that normatively support intentional aims and actions are *practical* reasons, and reasons that normatively support cognitive beliefs are *epistemic* reasons.

Focusing now on practical reasons, in the recent and contemporary philosophical literature following on from Bernard Williams, there is a crucial and very familiar—indeed, all-too-familiar—distinction between *internal* reasons and *external* reasons.[221] As I am understanding this distinction, internal reasons *do* belong to an agent's set of motivations, and external reasons do *not* belong to an agent's motivational set.

More precisely, however, *internalists* about practical reasons hold that reasons both motivate and also justify our choices or actions, where "justification" can be taken in

> either (i) the *strong sense* of the agent's having a sufficient rational ground that provides an obligation for her to choose or act, for example, an agent is in her office at 3:00 pm because she has promised to meet someone there then, and is strongly justified in so doing—she ought to be there then, other things being equal,
> or (ii) the *moderate sense* of the agent's having a legitimating rational ground that provides a permission for her to choose or act even if there is no obligation, for example, an agent has the day off and is working in her office anyhow, but then at 2:30 pm she suddenly decides to go to the movies for the rest of the afternoon and chill out, and is moderately justified in so doing—it is OK and a good thing for her to chill out for the rest of the afternoon, other things being equal,
> or (iii) the *weak sense* of the agent's having a rational opportunity for her to choose or act even if there is neither an obligation nor a permission, for example, an agent has promised to meet someone at her office at 3:00 pm, but then at 2:30 pm she suddenly decides to go to the movies for the rest of the afternoon and chill out, thereby blowing off her appointment, and is only weakly justified in so

> doing—it can be perfectly well understood why she goes to the movies, but not condoned, other things being equal.

According to reasons-internalism, *all* practical reasons are internal reasons. Internalists—for example, Hume—normally also hold *a desire-based theory* about the nature of justifying reasons.

By contrast, *externalists* about practical reasons hold that while all practical reasons justify our actions in any of the three senses of justification, nevertheless at least some and perhaps all practical reasons fail to motivate our actions. So some or all practical reasons are external reasons. Externalists—for example, contemporary Kantians like Korsgaard or later Parfit—normally also hold an objective-value-based theory of the nature of justifying reasons.

These two opposed positions of internalism and externalism about practical reasons may seem to exhaust the logical space of possible philosophical theses. But that is not correct. This is because Kant, on my reading, and I too, hold the intermediate view that while *all* practical reasons are both motivating and justifying, nevertheless *some* practical reasons are justifying but *not* motivating. That, in turn, seems plainly paradoxical. So how can a contemporary Kantian "externalistic internalism" about practical reasons intelligibly and non-paradoxically be the case?

The simple version of the answer to this question is that on the contemporary Kantian view of practical agency I am developing and defending:

> (i) all practical reasons whatsoever do indeed justify and have at least the potential or power to motivate the intentional agent, but
> (ii) in some act-contexts, some instrumental practical reasons that normally and other things being equal *would have motivated* the intentional agent, *do not actually motivate* the intentional agent in those contexts.

The slightly more complicated version of the same answer says that in all and only those act-contexts in which an instrumental practical reason that normally and other things being equal would have motivated an intentional action, actually does fail to motivate an intentional action, then this state of affairs rationally occurs (as opposed to its merely causally occurring, for example, if a bomb had gone off and killed the intentional agent right then and there, thereby removing the real possibility of the agent's being motivated by anything) precisely because the intentional agent in those contexts has a desire-overriding, strictly universal, a priori, categorically normative, non-instrumental practical reason that

> both (i) wholeheartedly motivates her to action in those contexts contrary to her egoistic or self-interested, hedonistic, or consequentialist inclinations,
> and also (ii) fully justifies her action in those contexts.

By "egoistic inclinations," I mean self-interested inclinations that are either narcissistic inclinations (driven by "self-love" or *Eigenliebe*) or selfish inclinations (driven by "self-inflation" or *Eigendünkel*). More generally, I distinguish between

> (i) egoistic narcissistic (that is, self-loving) desires,
> (ii) egoistic selfish (that is, self-inflating) desires, and

(iii) egoistic prudential (that is, *rationally* self-interested) desires.

Someone's deep interest in promoting the welfare of the other members of his own family can be rationally self-interested, hence prudent, but neither narcissistic nor selfish. By contrast, someone's gambling obsession, even if it alienates all his friends, destroys his marriage, and gets him fired from his job, can be selfish but neither narcissistic nor prudent. And again by contrast, someone's excessive vanity and narcissism can lead him to perform, again and again, extremely dangerous feats of physical exertion (free-style climbing on skyscrapers, etc.), even to the point of getting himself killed, but—assuming he has no dependents, loving partner, or loving family—this is neither selfish nor prudent. On the other hand, obviously, narcissism, selfishness, and prudence are all perfectly compatible with one another too, in one and the same person's character, choices, and actions. Indeed, the politicians that one loves to hate frequently exemplify just this compatible combination of egoistic traits, and each one of them to a very high degree. The current so-called President of the USA, Donald Trump, is an outstanding real-world example.

In any case, here is the more fully spelled-out and non-paradoxical formulation of the uniquely intermediate "externalistically internalist" Kantian view about practical reasons that I want to defend:

> While normally and other things being equal all practical reasons are both motivating and justifying, nevertheless in some contexts in which things are neither normal nor equal, *and* this is specifically due to an intentional agent's more or less wholehearted non-instrumental motivation, *then* at least some instrumental practical reasons are justifying but not motivating in those contexts.

This in turn is really possible because, as I am interpreting Kant's theory of practical rationality and practical agency, this theory includes an early version of "the hierarchical-desire model of the will" later re-discovered by Harry Frankfurt.[222] According to Frankfurt's model, the rational human will is capable of having "effective first-order desires," which are desires for this or that, that do or will or would move us all the way to action, and also "second-order desires," which are desires about first-order desires, for example, the desire to desire this or that. Among our second-order desires, in turn, are special ones known as "second-order volitions," which self-consciously determine just which effective first-order desires are to be the ones that do indeed move us all the way to action. But in the specifically *contemporary Kantian* version of the Frankfurtian model, however, there are, among the second-order volitions, certain ones that are characteristic of what Kant and Frankfurt alike call *freedom of the will*, that operate by not only self-consciously determining just which effective first-order desires do indeed move us all the way to action (which expresses Frankfurtian freedom of the will), but also (and this is the specifically Kantian supplement) can

> either (i) de-rail an occurrent first-order desire which would otherwise have motivated the agent to action,
> or else (ii) newly-and-spontaneously generate an effective first-order desire that substitutes itself for an occurrent first-order desire which would otherwise have motivated the agent to action.

On this contemporary Kantian/Frankfurtian hierarchical-desire model, then, *Willkür,* or the human power of choice/sensible will, is the faculty of effective first-order desires, or first-order volitions, and *Wille*, or practical reason (whether impure/instrumental or pure/non-instrumental), is the faculty of second-order volitions. The power of choice effectively desires ends or goals, and the satisfaction of desires produces pleasure or psychological happiness. Practical reason, in turn, recognizes the objective values of these ends or goals. When practical reason recognizes ends also as means for the production of happiness, it is instrumental. And when practical reason responds to ends for their own sake, it is non-instrumental. Practical reason is therefore a spontaneous responsiveness of the agent to reasons, inhering in the structure of the rational human will itself, that guides the agent to intentional action.

This, in turn, entails the denial of Henry Allison's well-known "Incorporation Thesis," according to which even instrumental motivation to action requires a separate act of non-desire-based practical reason, incorporating any first-order desire as a *maxim*, that is, a conceptual-intellectual act-policy or rational resolution.[223] The Incorporation Thesis implies that practical reason is always the intellectualist "master," and not the anti-intellectualist Humean "slave," of the passions. By sharp contrast, according to my contemporary Kantian/Frankfurtian view, practical reason is neither the intellectualist *master* of the passions (that is, a capacity essentially external to the passions), nor the anti-intellectualist Humean *slave* of the passions (that is, a capacity essentially reducible to the passions). Instead, practical reason is the non-intellectualist *guide* of the passions (that is, a capacity essentially internal to the passions, but not reducible to the passions), as a non-mechanical, spontaneous reasons-responsiveness inhering in the formal hierarchical-desire constitution of the rational human will itself, that expresses the agent's reasons-directed intentional ability to guide and control her own passions for the natural purposes of choice and action. Put in terms of my neo-Aristotelian theory of immanent structural properties, practical reason is *the activating immanent structure of the human will*.[224]

For Kant, the recognition of a desire-overriding non-instrumental reason depends on the objective value of the Categorical Imperative, the moral law. But recognition of the Categorical Imperative or moral law also triggers an innate affective disposition in rational human agents for having more or less wholehearted, higher-order, essentially non-conceptual, non-self-conscious, non-propositional, non-inferential desires to achieve self-transcendence with respect to their egoistic (whether self-loving/narcissistic or self-inflating/selfish), hedonistic, or consequentialist inclinations, by desiring to be moved by non-egoistic, non-hedonistic, non-consequentialist effective first-order desires *alone*. This in turn is what Kant calls the "change," "reversal," or "revolution" of the human heart or power of choice/sensible will, insofar as a person responds affectively and effectively to the moral law:

> If by a single and unalterable decision a human being reverses the supreme ground of his actions by which he was an evil human being (and thereby puts on a "new man"), he is, to this extent, by principle and attitude of mind, a subject receptive to the good; but he is a human being only in incessant laboring and becoming, i.e., he can hope … to find himself upon the good (though narrow) path of constant *progress* from bad to better. For him who penetrates to the intelligible ground of the heart (the ground of all the maxims of the power of choice/sensible will) …[225] this is the same as actually being a good human being … and to this extent the change can be considered a revolution. (*Rel* 6: 48)

In other words, in his all-things-considered moral theory circa 1792, in *Religion*, Kant defends what I will call a *higher-order affective innatism* about motivation by non-instrumental reasons.

Now it is extremely important to recognize that *sometimes* this innately specified and spontaneously generated, higher-order, wholehearted, essentially non-conceptual, non-self-conscious, non-propositional, non-inferential desire for self-transcendence is in fact *near-Satanically evil*. This would be exemplified in any case in which someone cold-bloodedly tries to kill, or seriously harm, someone else, *even though he knows that he is bound to be shot by the police for doing so*.[226] But also sometimes—namely, when it results from recognition of the Categorical Imperative/moral law—this innately specified and spontaneously generated more or less wholehearted higher-order desire for self-transcendence is *morally good and right*, and constitutes a revolution of the human heart. That morally good and right self-transcendence rarely happens in human affairs is fully acknowledged by Kant. But it *is* really possible, and as Kant firmly believed, and as I also firmly believe, it sometimes actually happens. Moral badness and moral evil happen all the time, all-too-obviously—but *only* insofar as revolutionary good and right happen, or at least *really can* happen, too.

These remarks raise the very difficult question of precisely how morally disvalued or wrong choice or action is *ever* possible according to a Kantian theory of practical agency. Here is the classical problem in a nutshell, as crisply formulated by Korsgaard:

> It is the essential nature of action that it has a certain metaphysical property—autonomy in Kant's argument, constitutional unity in Plato's. This explains why action must meet the normative standard: *it just isn't action* if it doesn't. But it also seems as if it explains it rather too well, for it seems to imply that only good action really is action, and that there is nothing left for bad action to be.[227]

This, in turn, has the direct implication that *the more morally evil or bad you are*, then *the less what you choose or do is authentically action*, and *the less you are morally responsible for choosing or doing it.* Or otherwise put, if moral responsibility requires freedom of the will, and morally wrong (whether evil or bad) choice or action is unfree because it is not autonomous, then human agents are *never* morally responsible for wrong choice or action. But that is a morally absurd result—or at the very least, it seems highly paradoxical.

My own view is that this absurd or at least seemingly highly paradoxical result is the immediate consequence of Korsgaard's mistakenly applying a One-Dimensional or 1D Conception of rational normativity to the Kantian theory of practical agency. By sharp contrast, the application of the Two-Dimensional or 2D Conception of rational normativity, together with affective innatism, to the Kantian theory of practical agency, effectively and smoothly solves the problem.

The basic idea behind the Kantian 2D solution to the problem is this. Every normal, healthy, human minded animal, as s/he matures, naturally and eventually comes to possess a set of basic online capacities for human rationality: then s/he is an agent in the minimal, nonideal, or "low bar" sense of meeting rational normative standards. This in turn is sufficient for causal and moral responsibility with respect to choices and actions, other things being equal. Then, morally disvalued or wrong choice or action is free choice or intentional performance that *meets* the minimal, nonideal, or "low bar" standards, yet also *falls short of* the maximal, ideal, or "high bar" standard of realized autonomy, occurrently having a good will, or occurrently acting for

the sake of the Categorical Imperative. Or in other words, morally wrong action is simply, to use another classical Augustinian notion, the "privation"[228] of maximally, ideally, or perfectly good and right action, while still remaining fully within the framework of the innate capacity for minimally, nonideally, or imperfectly good action. All morally wrong choice or action flows from what Kant calls the fully natural "propensity to evil in human nature" or "radical evil," that is, the fully natural and "human, all too human" disposition or tendency towards freely chosen moral disvalue or wrong, that is fundamentally caused by human egoism in all its forms, especially including self-love/narcissism (*Eigenliebe*) and self-inflation/selfishness (*Eigendünkel*). And this disposition or tendency is essentially the same as the other classical Augustinian notion I mentioned earlier in this section, "the perversity of the will,"[229] or what Kant calls "the perversity of the human heart" (*Rel* 6: 30, 37).

Now some morally wrong choice or action is just that: nothing but a privation of maximally, ideally, or perfectly good and right action. Such choice or action also occurs under what Joseph Raz aptly calls "the Guise of the Good,"[230] which is to say that it conforms to the classical Socratic idea that morally wrong choice or action is an erroneous, ignorant, or otherwise rationally misguided attempt to choose or do the good. That is *non-evil morally bad* choice or action. In turn, however, *evil* choice or action is a special sub-class of morally wrong choice or actions that involves the intention to undermine or violate the dignity of real persons. When evil choice or action *also* occurs under the Guise of the Good, and as it were Socratically, then it is *banal evil*. But when morally wrong choice or action is taken together with the innate capacity for having the desire for self-transcendence that is postulated by the Kantian higher-order affective innatism I spelled out just a few paragraphs back, then—as against Socrates, who thought this was impossible—the rational human animal or real human person simply chooses the wrong thing *for its own sake*, or non-instrumentally. This is what Raz also aptly calls choice or action under "the Guise of the Bad."[231]

Formulated explicitly and fully now, according to this contemporary Kantian account of the real possibility of morally disvalued or wrong choice or action, *then morally disvalued or wrong choice or action has three individually necessary and jointly sufficient features*, and two distinct sub-kinds (namely, moral evil and non-evil moral badness), one of which (namely, moral evil) itself has two further distinct sub-kinds. Or, fully explicitly now, a practical agent *A* chooses or acts in a morally disvalued or wong way if and only if

> (i) *A* satisfies the minimal, nonideal, or "low bar" standards of rational normativity, thereby guaranteeing moral responsibility, and also
> (ii) *A* falls short of the maximal, ideal, "high bar" standards of moral rational normativity, namely, the Categorical Imperative or moral law (= morally bad action as the "privation" of ideally or perfectly good and right action), and also
> (iii) *A* freely chooses or does the morally disvalued or wrong thing *insofar as this choice or action is perverted by human egoism in any of its forms*, thereby flowing from "the perversity of the will" or "the perversity of the heart," which, in turn, can be
> EITHER (iii.1) *moral evil*, that is, choice or action involving the intentional violation of people's dignity, which is the same as treating them like like mere instruments, or what is worse, like mere things, or, what is even worse, like mere garbage or offal, which, in turn, can be either (iii.1i) *near-Satanic evil*, that is, evil chosen or done for its own sake and under the Guise of the Bad, whatever the

> consequences, as the result of *titanic egoism*—for example, Hitler, or (iii.1ii) *banal evil*, that is, evil chosen or done for merely self-interested reasons (aka "banal egoism"), for hedonistic reasons, or for consequentialist reasons, and under the Guise of the Good —for example, (Arendt's) Eichmann,
> OR (iii.2) *non-evil moral badness*, which is the non-evil privation of high-bar good, for example, not doing enough to help others.

And that Kant himself is at least implicitly committed[232] to this ramified, subtle 2D Conception is clearly implied by these texts:

> CONCERNING THE PROPENSITY TO EVIL IN HUMAN NATURE. By *propensity* … I mean the subjective ground of the possibility of an inclination (habitual desire, *concupiscentia*), insofar as this possibility is contingent for humanity in general. It is distinguished from predisposition in that such a propensity can indeed be innate yet *may* be represented as not being such: it can rather be thought of (if it is good) as *acquired*, or (if evil) as *brought* by the human being *upon* himself. —Here, however, we are only talking of a propensity to genuine evil, i.e., moral evil, which, since it is only possible as the determination of a free power of choice and this power for its part can be judged good or evil only on the basis of its maxims, must reside in the subjective ground of the possibility of the deviation of the maxims from the moral law. And, if it is legitimate to assume that this propensity belongs to the human being universally (and hence to the character of the species), the propensity will be called a *natural* propensity of the human being to evil. —We can further add that the power of choice's capacity or incapacity arising from this natural propensity to adopt or not to adopt the moral law in its maxims can be called *the good or evil heart*. (*Rel* 6: 29, underlining added)

> According to our mode of estimation, [to us] who are unavoidably restricted to temporal conditions in our conceptions of the relationship of cause to effect, the deed, as a continuous advance *in infinitum* from a defective good to something better, always remains defective, so that we are bound to consider the good as it appears in us, i.e., according to the deed, as *at each instant* inadequate to a [moral] law. But because of the *disposition* from which it derives and transcends the senses, we can think of the infinite progression of the good toward conformity to the law as being judged by him who scrutinizes the heart … to be a perfected whole even with respect to the deed (the life conduct). (*Rel* 6: 67, underlining added)

If I am correct about all this, then the classical *Humean* and Kantian accounts of practical agency are *much* closer both in detail and spirit than has previously been thought. The crucial difference between them is Kant's idea that the motivational force of an overriding morally right practical reason can be based exclusively on an innate emotional disposition for having higher-order wholehearted desires to be moved by morally appropriate non-egoistic or non-self-interested, non-hedonistic, non-consequentialist first-order desires. This innate emotional disposition, which corresponds to what Kant calls the capacity for "respect" or *Achtung*, is causally triggered by a person's subjective experience or consciousness of recognizing the Categorical Imperative as an overriding, strictly universal, a priori, categorically normative,

non-instrumental practical reason. As I also mentioned above, this subjective experience or consciousness of recognizing the Categorical Imperative, in turn, is the Fact/Affect of Reason.

In order to explicate, develop, and then defend this contemporary Kantian theory of practical freedom as free agency, I want to look more closely now at Kant's *rational teleology*, that is, his theory of practical ends or purposes, and also at his corresponding theory of the constitution of the human will, with special reference to its internal psychological structure and its characteristic activities or operations. Here are the relevant texts.

> The will is a capacity to determine itself to acting in conformity with the *representation of certain laws*. And such a capacity can be found only in rational beings. Now, what serves the will as the objective ground of its self-determination is an end, and this, if it is given by reason alone, must hold equally for all rational beings. What, on the other hand, contains merely the ground of the possibility of an action the effect of which is an end is called a *means*. The subjective ground of desire is an *incentive*; the objective ground of volition is a *motive*; hence the distinction between subjective ends, which rest on incentives, and objective ends, which rest on motives, which hold for every rational being. Practical principles are *formal* if they abstract from all subjective ends, whereas they are *material* if they have put these, and consequently certain motives, at their basis. The ends that a rational being proposes at his discretion as *effects* of his actions (material ends) are all only relative; for only their mere relation to a specially constituted faculty of desire on the part of the subject gives them their worth, which can therefore furnish no universal principles, no principles valid and necessary for all rational beings and also for every volition, that is, no practical laws. Hence all these relative ends are only the ground of hypothetical imperatives. But suppose that there were something the *existence of which in itself* could be a ground of determinate laws; then in it, and in it alone, would lie the ground of a possible categorical imperative, that is, of a practical law…. Beings the existence of which rest on our will but on nature, if they are beings without reason, still have only relative worth, as means, and are therefore called *things*, whereas rational beings are called *persons* because their nature already marks them out as an end in itself, that is, as something that may not be used merely as a means, and hence so far limits all choice (and is an object of respect). These, therefore, are not merely subjective ends, the existence of which as an effect of our action has a worth *for us*, but rather *objective ends*, that is, beings the existence of which is in itself an end, and indeed one such that no other end, to which they would serve *merely* as a means, can be put in its place, since without it nothing of *absolute worth* would be found anywhere; but if all worth were conditional and therefore contingent, then no supreme practical principle for reason could be found anywhere. (*GMM* 4: 427-428)
>
> In the kingdom of ends everything has either a *price* or a *dignity*. What has a price can be replaced by something else as its *equivalent*; what on the other hand is raised above all price and therefore admits of no equivalent has a dignity. What is related to general human inclinations and needs has a *market price*; that which, without presupposing a need, conforms with a certain taste, that is, with a delight in the mere purposeless play of our mental powers, has a *fancy price*; but that which constitutes the condition under which alone something can be an end in itself has not merely a relative worth, that is, a price, but an inner worth, that is, a *dignity*. (*GMM* 4: 434-435)
>
> All material practical principles put the determining ground of the will in the *lower faculty of desire*, and were there no *merely formal* laws of the will sufficient to

> determine it, then neither could *any higher faculty of desire* be admitted…. The principle of one's own happiness, however much understanding and reason may be used in it, still contains no determining ground for the will other than such as is suitable to the *lower* faculty of desire…. Then, only insofar as reason of itself (not in the service of the inclinations) determines the will, is reason a true *higher* faculty of desire, to which the pathologically determinable is subordinate, and then only is reason really, and indeed *specifically*, distinct from the latter, so that even the least admixture of the latter's impulses infringes upon its strength and superiority. (*CPrR* 5: 22, 24-25)
>
> The capacity for desiring in accordance with concepts, insofar as the ground determining it to action lies within itself and not in its object, is called the capacity for *doing or refraining from doing as one pleases*. Insofar as it is joined with one's consciousness of the capacity to bring about one's object by one's action it is called *the capacity for choice* (*Willkür*); if it is not joined with this consciousness its act is called a *wish*. The capacity for desire whose inner determining ground, hence even what pleases it, lies within the subject's reason, is called the *will* (*Wille*). The will is therefore the capacity for desire considered not so much in relation to action (as the capacity for choice is) but rather in relation to the ground determining choice to action. The will, strictly speaking, has no determining ground; insofar as it can determine the capacity for choice, it is instead practical reason itself. Insofar as reason can determine the capacity for desire in general, not only *choice* but mere *wish* can be included under the will. The choice which can be determined by *pure reason* is called free choice. That which can be determined only by *inclination* (sensible impulse, *stimulus*) would be animal choice (*arbitrium brutum*). Human choice, however, is a capacity for choice that can indeed be *affected* but not *determined* by impulses, and is therefore of itself (apart from an acquired aptitude of reason) not pure but still can be determined to action by pure will. *Freedom* of choice is this independence from being *determined* by sensible impulses; this is the negative concept of freedom. The positive concept of freedom is that of the capacity of pure reason to be itself practical. But this is not possible except by the subjection of the maxim of every action to the condition of its qualifying as universal law. (*MM* 6: 213-214)

As Tamar Schapiro has correctly pointed out, desires in the Kantian sense are not merely preferences for something, pro-attitudes towards something, or wishes for something. On the contrary, above and beyond those less active and less committed mental states or acts, a desire in the Kantian sense is a *felt need* for something, a conscious *going-for* something.[233] According to Kant as I am reading him, and according to my contemporary Kantian view, then, desires are always felt needs or conscious goings-for, aimed at ends. Objective ends are intrinsic values, and provide motives for action—namely, the motives directly corresponding to our motivating desires. Subjective ends are the pleasurable satisfactions of desires and the removal (or anyhow the control) of painful frustrations of desires, and provide *incentives* for action. Means are things valued only for the sake of ends, hence are only extrinsic values. Objective ends can have either a *price* or a *dignity*. For an end to have a price means that it has some equivalent which can be substituted for it. Price can either be *market* price (in terms of satisfaction of interests) or *fancy* price (in terms of disinterested satisfaction).

Dignity, by a fundamental contrast, is *absolute intrinsic objective value*, which is a real and objective yet non-empirical mode of value that is beyond all denumerable evaluative quantity, or price. This is because price includes, as a necessary condition, that it is a human-scaled

(market- and fancy-) value quantity for which some denumerable equivalent, that is, a rational or natural number equivalent, can be fixed. Price, in other words, is an *economic* value quantity. Now all and only real ends-in-themselves, or real persons, have dignity. In turn, the value of real persons, and non-instrumental moral value generally, is essentially a transfinite quantity, and human economics does not deal in transfinite quantities. I am not saying that economists cannot or do not use real numbers or complex numbers in their calculations; rather, what I mean is that when they translate their calculations into real-world quantities—for example, of things bought, sold, and traded; or of dollars and cents—then *those* quantities are all denumerable. Correspondingly, contrary to popular belief, the "dismal science" of economics is not dismal because it is pessimistic: it is dismal *because it is not about anything that is intrinsic to the lives of real persons*. Hence the value of *every* real person, and *only* what partakes in the value of real persons, is inherently beyond all actual or possible human economics.

As we have already seen in passing, the Highest or Supreme Good of Kant's ethics is having a *good will* in Kant's sense (*GMM* 4: 393) (*CPrR* 5: 110-111), that is, occurrently partially or completely realizing autonomy. *The Realm of Ends*, also called "the ethical community" (*Rel* 6: 94), is the total ideal or real[234] moral community of rational human animals or real human persons, each of whom respects one another and themselves as creatures with dignity (absolute objective intrinsic non-denumerable moral value), and also considers all the others and themselves equally in relation to the Categorical Imperative/moral law; and, finally, each possesses a good will. The "sole and complete good," (*GMM* 4: 396)—that is, the best life for any rational human animal or real human person—is a life of deep individual happiness and also deep communal or social happiness, that is "deep" precisely insofar as as it is intrinsically controlled and structured by a good will in the Kantian sense, and not merely "shallow" or morally accidental happiness.

Here, in turn, is Kant's basic theory of the constitution of the human will, which I also fully accept—or at least I fully accept it *as* I am interpreting it, according to the Kantian/Frankfurtian hierarchical-desire model of the will I sketched above. The human will, or *faculty of desire* (*Begehrungsvermögen*), is our innate capacity for mobilizing and organizing our desires in order to motivate or move ourselves to choosing or doing, and in human persons the will is a rational human agent's power of wanting, intending, deliberating, deciding, and trying. In turn, the human will or the faculty of desire has two levels:

> (i) the lower or executive faculty of effective first-order desires, namely, first-order volitions, *the power of choice* (*Willkür*), and
> (ii) the higher or legislative faculty of second-order volitions, the will (*Wille*) proper, *or the faculty of practical reason.*

So the faculty of practical reason is a necessary proper part of the human will or faculty of desire. Moreover, the faculty of practical reason is the will in the proper or rational sense. In these ways, the faculty of practical reason is "the will proper" in two senses: it is an inherent part of the human will, and indeed the action-guiding core of the human will; and it also encapsulates the primary aim or end of the human will, which is to be a self-realizing rational human animal—that is, to be a creature that is principled and authentic, at least to some salient degree and extent. Otherwise put, according to my Kantian/Frankfurtian account, the human will or faculty of desire is a complex, integrated psychological structure that fuses desiderative animality and choice-or-action-guiding-and-controlling rationality,[235] and looks like this:

Human Will or Faculty of Desire (*Begehrungsvermögen)*
higher part = faculty of practical reason or will proper (*Wille)*
higher part = pure or non-instrumental reason
lower part = impure or instrumental reason
lower part = power of choice (*Willkür)*

In this Kantian/Frankfurtian model of the will, *Willkür*, or the power of choice, is an executive first-order volitional power of intentional causation by means of effective first-order desires. By contrast, *Wille*, or the will proper, is a higher-order volitional power of self-legislation that operates by means of recognizing either instrumental or non-instrumental reasons for the determination of choice. To act on the basis of *Willkür* is to move our animal bodies by means of our effective first-order desires. This can of course occur in a more or less Humean way by means of instrumental reasoning according to the hypothetical imperative. Since instrumental reasoning is itself a form of self-legislation, it thereby involves what we can call the "impure" *Wille*. Now the lower or executive faculty of desire, namely, the power of choice, is normally motivated or moved by objective ends that are picked out by our egoistic or self-interested, hedonistic, or consequentialist desires, and constitute the "matter" of our happiness, which is the pleasurable satisfaction of desires and the removal (or anyhow the control) of their painful frustration. Insofar as the faculty of practical reason is concerned with these ends, it is an "impure" and instrumental reason. This is the lower faculty of practical reason. But it is also possible for the faculty of practical reason to be pure and non-instrumental, and therefore to be moved not by the matter of our happiness, but rather solely by *the form of law-giving*, that is, by the structure of personhood or free agency itself, our essential nature as rational human animals, considered as an objective but purely formal end. This is the higher faculty of practical reason. The law that is given to themselves by human persons, that is, by free agents who are rational human animals, is the Categorical Imperative or moral law, hence higher willing of this type is positive freedom or realized autonomy.

3.4 THE RATIONALITY OF THE HEART: PRINCIPLED AUTHENTICITY

To act on the basis of our "pure" *Wille*, our own capacity for pure practical reason, however, is to constrain and determine our *Willkür*, our own power of choice, by recognizing the Categorical Imperative. Insofar as it is recognized by us, the Categorical Imperative provides a desire-overriding, strictly universal, a priori, non-instrumental reason for action. That recognition, in turn, causally triggers an innate higher-order emotional disposition in us (also known as the capacity for *Achtung* or respect) to generate the consciously experienced desire to be moved by morally appropriate and non-egoistic or non-self-interested, non-hedonistic, non-consequentialist effective first-order desires:

> The direct determination of the will by the law, and the awareness of that determination, is called "respect," so we should see respect as the *effect* of the law on a person rather than as what *produces* the law. Actually, respect is the thought of something of such worth that it breaches my self-love…. Any moral so-called *interest* consists solely in *respect* for the [moral] law. (*GMM* 4: 402 n.)

Respect, in turn, is essentially a capacity and (when the capacity is triggered) an operation of the rationally-responsive *human heart*:

> The pure thought of duty and in general of the moral law, mixed with no foreign addition of empirical inducements, has by way of reason alone (which with this first becomes aware that it can of itself be practical) an influence on the human heart so much more powerful than other incentives, which may be summoned from the empirical field, that reason, in the consciousness of its dignity, despises the latter and can gradually become their master. (*GMM* 4: 410-411, underlining added)
>
> We betray a culpable degree of moral unbelief if we do not grant sufficient authority to duty's precepts, as originally inscribed in the heart by reason. (*Rel* 6: 84, underlining added)

> Moral faith (*Glaube*) must be a free faith, founded on pure dispositions of the heart (*fides ingenua*). (*Rel* 6: 115, underlining added)

> The highest goal of the moral perfection of finite creatures, never completely attainable by human beings, is … the love of the [moral] law. (*Rel* 6: 145, underlining added)

So to choose or act on the basis of pure *Wille* is to do the good and right thing, as determined by our own pure practical reason, via the unique, wholly heartfelt motivational influence of the innate dispositional higher-order emotion of respect on our effective first-order desires or choices, no matter what the external and psychological antecedents, no matter how much pain I might suffer by doing the right thing, and no matter what the consequences.

The crucial factor in this account is Kant's truly important idea, also fully endorsed by me, that there exists an innate emotional disposition in all rational human agents to experience spontaneously generated wholehearted higher-order desires to be moved by non-egoistic or non-self-interested, non-hedonistic, non-consequentialist effective first-order desires or choices. As I mentioned above, I call this special wholehearted higher-order desire *the desire to achieve self-transcendence* because it is a person-unifying, life-changing desire to achieve a radical volitional distancing with respect to our own egoistic or self-interested, hedonistic, or consequentialist first-order desires, and thus to be able to overcome the almost irresistible centripetal forces of the Dear Self and the Bottom Line. Now non-egoistic or non-self-interested, non-hedonistic, non-consequentialist first-order desires take the following general form:

> I want (not-) X, no matter how much egoistic or hedonistic unhappiness and pain I may experience in getting (not-) *X*, and no matter what the consequences.

So, correspondingly, the higher-order desire to achieve self-transcendence takes the following general form:

> I want (not) to want (not-) *X*, no matter how much egoistic or hedonistic unhappiness and pain I may experience in getting (not-) *X*, and no matter what the consequences.

Here, in turn, are two key points about the desire to achieve self-transcendence.

First, postulating the desire to achieve self-transcendence as the motivational ground of chooing or acting for the sake of the Categorical Imperative or moral law is *not* "emotional extremism" according to any specifically negative philosophical or moral connotation of that phrase.[236] On the contrary, it is perfectly rational and human—indeed a proper part of the authentic self-realization of practical human rationality—to be prepared to go to the wall for the things that really and truly matter most to you. It is not that you specifically *want* to go to the wall or that you are specifically *trying* to go to the wall. Indeed, it is a necessary part of the desire to achieve self-transcendence when it is specifically triggered by recognition of the Categorical Imperative, that *were there any live or relevant option short of that*, which also sustained the Highest or Supreme good, and also brought about good consequences for others and for oneself, *then you would go for that instead*. It is just that the Highest or Supreme good is *worth more* than any actual or possible set of good consequences alone. So good consequences do not necessarily determine the Highest or Supreme good.

It is important to note that trying to bring about good consequences for others and oneself, other things being equal, is a strict moral obligation, and hence good consequences for others and oneself constitute a proper part of the Highest or Supreme Good (= a good will = a life of principled authenticity, achieved at least partially or to some degree), according to the sketch of contemporary Kantian ethics I am developing here, and also work out in full detail and defend in *Kantian Ethics and Human Existence*. The crucial thing to note here and now, however, is just that any attempt to substitute *that proper part* of the Highest or Supreme Good (= good consequences for others and oneself) for *the whole* of the Highest or Supreme Good (= a good will = a life of principled authenticity, achieved at least partially or to some degree) is fallacious.[237] The whole Highest or Supreme good, taken by itself, is worth nondenumerably infinitely more than any of its proper parts, and has its value whatever the consequences.

The second key point is that sometimes the desire to achieve self-transcendence is near-Satanically evil, and thus undertaken under the Guise of the Bad, as in the previously-mentioned case in which someone cold-bloodedly tries to murder or seriously harm someone else, *purely for the sake of violating (respect for) their dignity*, even though he knows that he is bound to be shot by the police. In such a case, the intrinsic value, or objective end, that triggers the higher-order desire to achieve self-transcendence is the fact that the assailant chooses and acts on the morally disvalued and wrong goal and reason, no matter how much pain he may experience in bringing it about, and no matter what the consequences.

As I say, this is a clear and distinct case of near-Satanic evil, and I have already spelled a working analysis of the conditions of its real possibility, within the framework of The 2D Conception of rational normativity. Here I want to stress that near-Satanic evil minimally implies our ability to act with transcendental freedom of the will, namely, agentive sourcehood or up-to-me-ness, but also in a monstrously self-loving/narcissistic and self-inflating/selfish way, and wrongly, hence without practical freedom of the will, a good will, or occurrently realized autonomy. Of course it must also be added that both the capacity for and also the realization of transcendental freedom entail our possession of the capacity for practical freedom, since all of these can occur only in a rational intentional agent, and amongst the animal intentional agents, only (as far as we know, leaving aside possible rational alien intentional agents) in a rational human intentional agent (*CPR* A533-534/B561-562). But near-Satanic evil

also implies our ability to act freely on the basis of innately-specified and spontaneously generated, highly maleficent, but also non-egoistic, non-hedonistic, or non-consequentialist *first-order* desires. It is possible wholeheartedly to want a thing that violates human dignity, no matter how much first-order egoistic and hedonistic unhappiness and pain you experience in getting it, and no matter what harmful things to you or anyone else as a consequence of your actions. So you want that evil thing for its own sake, literally *for the hell of it*. In this way, just like Hume, Kant does not regard it as contrary to reason (in the low-bar or nonideal sense of 2D rational normativity), for someone to prefer the destruction of the world, including his own self-destruction, to the scratching of his finger.[238]

In order to make this rational human emotional possibility of near-Satanic evil even more concrete, we can think not only of Hume's monstrously narcissistic/self-loving and selfish/self-inflating person, but also of John Milton's Satan, a literary character that Kant in all likelihood would have known well,[239] of Iago in Shakespeare's *Othello*; of the nihilist Peter Verhkovensky in Dostoevsky's *The Devils*; of the post-modern punk thug Alex in Anthony Burgess's/Stanley Kubrick's *Clockwork Orange*; of the brilliant serial-killer/cannibal Hannibal Lecter in Jonathan Demme's *Silence of the Lambs*; of the ultra-cold-blooded hitman Anton Chigurh in Cormack McCarthy's novel/the Coen Brothers' movie *No Country for Old Men*; of the real-life totalitarian mass-murderers, Hitler and Stalin; of the real-life perpetrators of the Columbine High School Massacre in 1999, Eric Harris and Dylan Klebold; of the homegrown US terrorist Timothy McVeigh, federally executed in 2001 for the Oklahoma City Bombing in 1995; and, sadly, of many, many other moral monsters, too numerous to mention.

Such evil, although it is monstrously egoistic, and near-Satanic, is also low-bar and nonideally *rational*. Only a *rational*, even though "human, all too human," animal could ever have such a self-transcending desire. Indeed, on Kant's account of the nature of desire, no desires had by rational human animals could ever be essentially *irrational* or *arational*, since the function of a desire in a rational human animals is just to move themselves to action in the service of attaining rationally-recognized objectively intrinsically valuable instrumental or non-instrumental ends—whether these are material ends, in the case of empirical desires based on pleasure and pain, or formal ends, in the case of moral emotion of respect (*CPrR* 5: 21-28). But some non-egoistic, non-hedonistic, non-consequentialist desires are more rational than others, and some are monstrously egoistic. So according to Kant, it would be monstrously egoistic for me to prefer the destruction of the world (including my own self-destruction) to the scratching of my finger, precisely because this would be a radical violation of the Formula of Humanity as an End in Itself version of the Categorical Imperative:

> So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means. (*GMM* 4: 429)

I would thereby be considering and treating everyone in the world, including myself, as mere things and mere means to my own ends, and worth less than my momentary mild pain, for its own categorically wicked sake and no matter what.

"For its own categorically wicked sake, and no matter what." Therefore, in an apparent paradox, near-Satanic evil is *the exact reversed mirror-image of acting with a good will*, for good ends, without self-interest, non-hedonically, for its own categorically good sake, whatever the consequences—the "flip side" of a Kantian moral saint. Correspondingly, the other seemingly paradoxical thing about near-Satanic evil, as per Kant's profound idea of a

"revolution of the heart" or "revolution of the will," and as per Dostoevsky's novels, given the transcendentally free, low-bar rational, double-edged potentiality of the desire for self-transcendence, is that it only takes *one* life-changing, existential Kantian act of choice to go over completely from Great Sinner to Saint.

Now, in your moral first-person imagination, starting out as a Great Sinner or as a banal sinner, as a Hitler/Stalin or as an (Arendtian) Eichmann, or even just as an everyday non-evil, morally bad person, now *revolutionize your will*, and if you are evil, then *flip the flip side*. Therefore the desire for self-transcendence, whenever it results from recognition of the Categorical Imperative/moral law, thereby motivating choice or action for the sake of the Categorical Imperative/moral law, is inherently morally good and right. That inherently morally good and right self-transcendence only *rarely* happens in human affairs is fully acknowledged by Kant: "from the crooked timber of humanity, nothing straight can ever be made" (*IUH* 8: 23). But on the other hand, Kant firmly believed, even given the inherent crookedness of our timber, our sheer bent humanity, it *is* really possible.

In support of this, he provides a famous thought-experiment of a very lustful person who (sharply unlike the near-Satanic and monstrously egoistic Humean person, who prefers the destruction of the world, including his own self-destruction, to the scratching of his finger) would never in fact gratify his lust and thereby commit a crime for any instrumental reason, if at the moment of committing his crime he were presented with the gallows from which he would be instantly strung up as punishment. But this very same very lustful person nevertheless regards it as fully possible for him to lay down his own life on the very same gallows by refusing to give false testimony against an honorable man, even though he were commanded to do on pain of death by a tyrannical prince, and thus the very lustful person conceives it to be really possible for him to choose and act on the basis of a moral non-instrumental reason (*CPrR* 5: 30).

But how is this radical "change," "reversal," or "revolution" of the human heart, that is, of the power of choice/sensible will, really possible, for a very lustful or even near-satanically evil person? The quick Kantian answer is that it *can* happen because it *ought* to happen, *and* because we all actually *do* innately have the power (for transcendental freedom, practical freedom, and autonomy) to make it happen. The longer Kantian answer is that it results from our spontaneous power of choice together with the intense personal struggle to reconfigure the basic hierarchical-desire-structure of one's own embodied will so as to achieve increasingly greater degrees of moral authenticity and self-transcendence, thereby overcoming egoism. How few of us ever manage to do this! And even for those few who do, how infrequently it ever happens! But in any case, frankly, I also think it actually *does* happen sometimes, and *more often* than you might think.

To be sure, publicly-acknowledged examples of moral heroism, sainthood, or radical goodness and rightness in the Kantian sense are not, to put it mildly, as thick on the ground as the relentlessly repetitive, Wheel of Ixion-like, trillionfold examples of near-Satanic or banal evil, and moral badness. Moreover, many or even most examples of moral heroism, sainthood, or radical goodness and rightness in the Kantian sense consist in all-too-tempting evil things *not* chosen and *not* done. Consider this imaginary headline—

> Woman Resists Terrible Temptation to Do Evil; Intensely But Quietly Suffers For It; Experiences Moral Self-Fulfillment; Then Goes About Her Daily Life As Usual.

Such stories rarely make it into *USA Today* or onto Fox News, nor even into *The New York Times* or *WaPo*. I mean, of course, that they never do and they never will. And the representation of radical goodness and rightness in art, sadly, also tends to be moralistic and prissy.

But in order to make the rational human emotional possibility of radical goodness and rightness more concrete, we can think here of Socrates as represented by Plato in the *Apology*, the *Crito*, and the *Phaedo*; of the absurd "Knight of the Sorrowful Countenance," Don Quixote, in Cervantes's *Don Quixote*; of Kierkegaard's "Knight of Faith" in *Fear and Trembling*; of the "Idiot" Prince Myshkin in Dostoevsky's *The Idiot*; of Renée Falconetti's brilliant portrayal of Joan of Arc in Carl Theodor Dreyer's *Passion of Joan of Arc*; of Takashi Shimura's equally brilliant portrayal of the dying civil servant Kanji Watanabe in Kurosawa's *Ikiru*; and also of the real-life, and therefore "human, all too human," but still genuine moral heroes Abraham Lincoln, Mahatma Ghandi, Martin Luther King Jr., and Mother Teresa. And leaving aside these famous real-world cases, I also strongly believe *that there are some people you actually know, or at least have met, or have heard about, who, in their own (perhaps) less consistent and (certainly) less conspicuous ways, are relevantly like this.* And if *they* can be like that, and choose and act like that, then *you* can too.

So according to Kant, the real possibility of radical human goodness and rightness via a radical change, reversal, or revolution of the human heart is just an actual fact, although of course a unique sort of fact, namely a non-empirical fact about the rational human heart. More precisely, however, this unique non-empirical fact is the fact that our subjective experience or consciousness of recognizing the Categorical Imperative/moral law triggers our innate higher-order emotional disposition for feeling respect, that is, our love for the moral law that is inscribed in the human heart, and then respect spontaneously generates the wholehearted higher-order desire to achieve morally right self-transcendence: *straight from the heart*. So it is not only a non-empirical fact, but also an inherently affective, essentially non-conceptual, non-self-conscious, non-propositional, and non-inferential fact. This subjective experience, or consciousness, of our direct recognition of the Categorical Imperative/moral law, together with its higher-order, essentially non-conceptual, non-self-conscious, non-propositional, and non-inferential emotive causal-generative effects, is nothing more and nothing less than The Fact of Reason, and the text in which Kant introduces it bears repeating:

> The consciousness of this fundamental law [of pure practical reason, which says: so act that the maxim of your will could always hold at the same time as a principle of universal law giving] may be called a fact of reason, since one cannot ferret it out from antecedent data of reason, such as the consciousness of freedom (for this is not antecedently given), and since it forces itself upon us as a synthetic proposition a priori based on no pure or empirical intuition… In order to regard this law without any misinterpretation as given, one must note that it is not an empirical fact, <u>but the sole fact of pure reason, which by it proclaims itself as originating law</u>. (*CPrR* 5: 31, underlining added; see also 5: 6, 42-43, 47-48, 55-57, 91-94, and 104-108)

Strictly speaking, Kant often refers to this as *a* fact of reason—although of course in this text he does refer to it as "the sole fact of pure reason"—but I will continue to refer to it as *The* Fact of Reason, simply in order to distinguish it from other actual or possible facts of reason that may not be strictly speaking *moral* facts of reason, such as the regulative use of ideas of

pure reason in natural scientific inquiry (*CPR* A642-652/B670-680) and the closely-related regulative use of ideas of pure reason in biology and teleological contexts more generally (*CPJ* 5: 167-198).

It should be especially noted that, as I am construing it, The Fact of Reason is *perfectly consistent with* this famous text in the *Groundwork*:

> In fact, it is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action otherwise in conformity with duty rested simply on moral grounds and on the representation of one's duty…. (*GMM* 4:407)

This famous text from the *Groundwork* has always—or almost always—been taken to imply a strong and specifically Kantian version of moral epistemic skepticism, by saying that we cannot ever know with certainty whether we are acting or have acted *for the sake of* the Categorical Imperative, that is, *from duty*, and not merely *in conformity with duty*, but actually from egoistic or hedonistic or utilitarian motives. But my reading of this famous text is sharply different. I think that the crucial sub-phrase in Kant's key phrase, "it is absolutely impossible by means of experience to make out with complete certainty a single case…," is "*by means of experience*." More precisely, then, I think that it is perfectly consistent with what Kant has actually written at *Groundwork* 4: 407, to claim that, *yes*, *empirical knowledge*, in Kant's technical senses of "empirical" and "knowledge" (*Wissen*) of whether we are acting or have acted from duty or not, is impossible. Empirical knowledge for Kant necessarily involves empirical concepts and empirical intuitions, and has its meaning, justification, and truth necessarily determined, at least in part, by sensory, contingent facts about the natural world and ourselves. Moreover, at most, empirical knowledge can have *empirical certainty*, which is when a subject "takes something [namely, a judgment or proposition] to be true" (*Fürwahrhalten*), in such a way that this taking-to-be-true has a kind of indubitability or self-evidence which is both subjectively sufficient, which makes it "conviction" (*Überzeugung*) and also intersubjectively or objectively sufficient (*CPR* A820-822/B848-850), which entails that the judgment or proposition is also actually true. Examples of empirical certainty would be cases of ordinary perceptual judgments under highly favorable contextual conditions, such as G.E. Moore's famous anti-skeptical thesis, "This is my right hand and this is my left hand."[240] I think that Kant is absolutely correct that we cannot ever have knowledge *in this sense* of whether we are acting or have acted from duty or not.

Nevertheless, even while conceding that empirical knowledge of the morality of our own actions is impossible, we can *also* consistently and also justifiably claim on Kantian grounds that we still have the capacity for *veridical, direct, occurrent awareness* of our choosing and acting from from duty, by means of *non-empirical, essentially non-conceptual, moral self-consciousness.* In other words, even though we cannot have empirical knowledge of whether we are acting or have ever acted from duty or not, we can still have *non-empirical certainty* about this. Correspondingly, I also think that The Fact of Reason is precisely this veridical, direct, occurrent, non-empirical, essentially non-conceptual, moral self-consciousness of our acting from duty. Therefore, given The Fact of Reason, then we can indeed have veridical, direct, occurrent, essentially non-conceptual, moral self-consciousness, with non-empirical certainty, when (and only when) we are acting for the sake of the Categorical Imperative and

from duty. Otherwise put, acting from duty *is partially constituted by its own self-validating phenomenology*.

In any case, what is most crucial here is to note, again, is that The Fact of Reason is not a conceptual, self-conscious, propositional, inferential, or more broadly speaking *intellectual* fact, but instead an inherently affective or heartfelt, essentially non-conceptual, non-self-conscious, non-propositional, non-inferential, or more broadly speaking *affective* fact about how the higher-order moral feeling of respect, that is, love for the moral law inscribed in the human heart, operates on the hierarchical-desire-structure of the human will. So as I noted above, The Fact of Reason is really The Affect of Reason. All rational facts are either absolutely (that is, without requiring empirical inputs) or relatively (that is, requiring empirical inputs) spontaneously active; and the Fact/Affect of Reason is absolutely spontaneously active, insofar as it is absolutely spontaneously *passionate*. In other words, The Fact/Affect of Reason is a rational act of the human heart, not merely a rational act of the head. In this respect, Kant's view is strikingly like that of Pascal, who famously wrote that the human heart has (RH, on behalf of Kant: pure practical or moral) reasons of its own that (RH, on behalf of Kant: theoretical or scientific) reason knows nothing about.[241]

For example, someone raises her arm and shrieks in order to stop a street crime, or perhaps she becomes a whistleblower in a corporate, governmental, or scientific organization,[242] just because she feels in her heart and in her non-conceptual mind that it is the morally right thing to do, even though she thereby risks her own life (in the case of stopping the street crime), or even though she risks losing her job and all her co-worker friends, and perhaps also ruining her career and her marriage (in the case of the corporate, governmental, or scientific whistleblower), and even though she desperately wants to avoid "being involved." It seems clear that given these background conditions, only a second-order volition driven by the innate affective capacity for respect, that is, a radical kind of love, could motivate such acts. Therefore she is doing her duty. According to Kant, "duty is the necessity of an action [done] from respect for the moral law" (*GMM* 4: 400). This says, as I am understanding it, that duty is the obligation that is binding on any act which is such that only the feeling of respect for persons and the Categorical Imperative/moral law innately inscribed in the human heart will suffice to move us, no matter what our first-order desires might happen to be, and no matter what the consequences.

In turn, there seem to be two importantly distinct ways in which the feeling of respect can move us by way of the second-order volition consisting of the more-or-less wholehearted desire to achieve a morally principled self-transcendence.

According to the first way, the higher-order wholehearted desire to achieve morally principled self-transcendence can take a particular online egoistic, hedonistic, or merely consequentialistic would-be effective first-order desire offline, and substitute a morally appropriate pre-existing or latent (but as yet non-effective) non-egoistic, non-hedonistic, non-consequentialistic (hence non-instrumental) first-order desire in its place, so that it becomes the effective one. To borrow Kant's example, the very lustful person can take his intense online first-order desire to avoid being hanged offline, and then substitute a pre-existing or latent first-order desire to avoid bearing false witness against an honorable man, so that this latter desire now becomes his first-order volition.

And according to the second way, assuming a complete set of selfish, egoistic, hedonistic, or consequentialist (hence instrumental) online first-order desires, together with another complete set of non-egoistic, non-hedonistic, or non-consequentialist (hence non-instrumental)

first-order desires, from which none has yet emerged as a would-be effective first-order desire, the higher-order wholehearted desire to achieve principled self-transcendence can spontaneously re-organize the emotional constituents of that state so as to produce a new online non-egoistic, non-hedonistic, or non-consequentialist (hence non-instrumental) effective first-order desire that is also morally appropriate. To borrow another of Kant's examples, a person who is by nature somewhat cold and unsympathetic towards other people, and furthermore has many troubles of his own, can nevertheless wholeheartedly generate a new effective first-order desire to be kind to someone else (*GMM* 4: 398-399). This sort of emotionally and practically generative absolute spontaneity is strictly analogous to the intellectually and theoretically generative absolute spontaneity that yields pure a priori knowledge.[243]

Now, again, I am not saying, nor is Kant saying, that this is an easy thing to do, nor am I saying, nor is Kant saying, that it happens very frequently in the course of everyday, personal or public life. But in fact, it actually happens a lot more often than you might think, as Rebecca Solnit's 2009 study of spontaneously-formed, altruistic "disaster-communities," *A Paradise Built in Hell*, compellingly shows. In any case, I think it is clearly really possible, and, just as clearly, sometimes actual. Indeed, as I indicated above, I also think that to some salient degree, *we have all either directly experienced this disposition or propensity in ourselves or else clearly recognized it in others*. At the very least, we have all recognized that we are fundamentally *capable* of it, because we have all recognized, no matter how fleetingly or darkly, *that a real human person can and should change her life for the better*. For example, one of the best-known lines in modern poetry is the last line of Rilke's "Archaic Torso of Apollo": *Du mußt dein Leben ändern*—"You must change your life."[244] Rilke's poetic imperative is clearly a Kantian imperative. Now, insofar as we can read Rilke's poem, and understand this line, then we morally *imagine* changing our own lives for the better, and rationally *hope* that we can.

We should also *not* assume, however, that the deeply motivational, desire-overriding, innate emotional disposition for feeling respect for persons and the moral law, or having the wholehearted desire to achieve morally principled self-transcendence, when triggered into action, will always have *the same phenomenology*. It may manifest itself as intense feelings of depression or gloominess (for example, Lincoln); of guilt; of sympathy; of empathy; of ecstatic suffering (for example, Falconetti's Joan of Arc); of intense outwardly-directed anger; or even of self-loathing (for example, Prince Myshkin). As Kant points out, since it "breaches my self-love" or narcissism, and also "humiliates," "strikes down," and "weakens" my self-inflation or selfishness, the subjective experience of respect or the desire for morally principled self-transcendence is often extremely unpleasant.

Of course, other things being equal, it is not terribly enjoyable, indeed "not my idea of a good time," to thwart one's own powerful egoistic, hedonistic, or consequentialist first-order desires. Freudians would call it *repression*, and this also raises a corresponding philosophical worry: If having a good will in the Kantian sense often involves repression, hence what is "not my idea of a good time," then how can it still be *good*? In reply to the Freudians, Kant could say:

> "Yes, I agree completely that it is repression, and also that repression, other things being equal, is "not my idea of a good time": not a happy experience. But precisely because we are crooked timbers and radically evil, a certain amount of serious repression is just the psychic cost of moral goodness, rightness, and virtue. The vitally important point is *that it is more-or-less wholehearted*, not *that it is normally either*

> *pleasant or self-gratifying*. Acting with a good will involves psychic coherence and self-sufficiency, not necessarily ordinary happiness."

This vitally important point requires more elaboration because it gets to the heart of Kantian non-intellectualism. There is repression, and then there is repression. Certainly, much repression is intensely unpleasant, morally and personally pointless, and even positively harmful. Think, for example, of all the thoroughly messed-up central characters in Hitchcock's most brilliant films from the notoriously repressed 1950s: *Strangers on a Train*, *Vertigo*, *Rear Window*, and especially (at the very end of the decade) *Psycho*. But a transcendentally (or in my terminology, deeply) free and rational human animal—that is, a conscious, self-conscious, and self-reflective human agent, capable of theoretical and logical a priori cognition, who also has the innate emotional and practical capacity for being motivated or moved by respect for persons and the moral law, or by the wholehearted desire to achieve morally principled self-transcendence—may *sometimes* be, but does not ever *have to* be, helplessly handcuffed, manipulated, overwhelmed, twisted, or violated by her own desires. This is because the innate emotional disposition for feeling respect for persons and the moral law, or the desire to achieve principled self-transcendence, essentially affectively expresses her deepest self.

Think again here of Falconetti's Joan of Arc, and of Mother Teresa, and of all those thousands or even millions of obscure, inconspicuous, unsung, perhaps only part-time, everyday moral saints that everyone actually knows, has met, or at least has heard about. If an agent is ever truly motivated or moved by respect for persons and the moral law, or by the wholehearted desire to achieve principled self-transcendence, even if it requires a terrible struggle to achieve it, then ultimately she has the will that she most deeply wants. She has thereby realized the capacity for rational emotional control of her own conscious, affective, and practical life. The internal constitution of the person she is and the person she will become are then both ultimately *up to her*: they flow from her, as the agential source. She is therefore both transcendentally (aka deeply) free and also practically free, hence partially or completely occurrently autonomous.

As I have noted already, Kant very aptly calls the subjective experience, or consciousness, of this special sort of moral self-control and agentive sourcehood, "self-fulfillment" or *Selbstzufriedenheit*:[245]

> Have we not, however, a word that does not denote enjoyment, as the word happiness does, but that nevertheless indicates a satisfaction with one's existence, an analogue of happiness that must necessarily accompany consciousness of virtue? Yes! This word is *self-fulfillment*, which in its strict meaning always designates only a negative satisfaction with one's existence, in which one is conscious of needing nothing. Freedom, and the consciousness of freedom as an ability to follow the moral law with an unyielding disposition, is *independence from the inclinations*, at least as motives determining (if not as *affecting*) our desire, and so far as I am conscious of this freedom in following my moral maxims, it is the sole source of an unchangeable fulfillment, necessarily combined with it. (*CPrR* 5: 117)

Such a state of rational, volitional self-fulfillment is a higher, and indeed a *higher-order*, kind of happiness that is analogous in certain respects to ordinary or first-order happiness, in that it results from the satisfaction of desires, but also sharply different in that it consists in the

satisfaction of a *special higher-order desire*—the desire for self-transcendence—not in the satisfaction of first-order desires. Moreover, it is essentially *deeper* than ordinary or first-order happiness. As "a negative satisfaction with one's existence (*ein negatives Wohlgefallen seiner Existenz*), in which one is conscious of needing nothing," it is the emotional state of rational, volitional coherence and self-sufficiency in a contingent, complex, and thoroughly nonideal actual world. Or to quote completely a Kant-text that I quoted partially above:

> *[c]onsidered in themselves* natural inclinations are *good*, i.e., not reprehensible, and to want to extirpate them would not only be futile but harmful and blameworthy as well; we must rather constrain them, so that they will not wear each other out but will instead be harmonized into a whole called happiness. (*Rel* 6: 58, underlining added).

Higher, higher-order happiness as *Selbstzufriedenheit* is Kant's anticipation of what the Existentialists later called "authenticity," or *Eigentlichkeit*, including what Kierkegaard called "purity of heart"—

> *Purity of Heart Is to Will One Thing* …. [T]he person who in truth wills only one thing *can will only the good*, and the person who wills only one thing when he wills the good *can will only the good in truth*,[246]

and also of what Frankfurt calls the "decisive identification" of second-order volitions with effective first-order desires or first-order volitions,[247]and when it is dynamically spread out over time, "wholeheartedness"—although, of course, in the present context with an inherent and specifically Kantian orientation towards respect for real persons and the Categorical Imperative/moral law. As I have said, I call this essential fact of free agency *principled authenticity*. Whatever we call it, along with Kant I do think it is a variety of free will most definitely worth having: indeed, Kant in my opinion altogether rightly thinks that it is the variety of free will *most* worth having.

One way of vividly highlighting the centrality of *Selbstzufriedenheit* to Kantian non-intellectualism, is to contrast it with its moral contrary, which is half-heartedness, impurity of heart, lack of heart ("my heart just wasn't in it"), or failure of heart—the various modes of inauthenticity in the specifically Kantian sense. This moral-psychological phenomenon of psychic incoherence and self-insufficiency appears in Kant's writings in at least three slightly different versions.

The first is the almost shockingly stark picture of the philosopher who dogmatically, Scholastically, and slavishly accepts the precepts of some existing philosophical system such as the Leibnizian-Wolffian philosophy:

> He has formed himself according to an alien reason, but the faculty of imitation is not that of generation, i.e., the cognition did not arise **from** reason in him, and although objectively it was certainly a rational cognition, subjectively it is still merely historical. He has grasped and preserved well, i.e., he has learned, and is a plaster cast of a living human being. Rational cognitions that are objectively so (i.e., could have arisen originally only out of the reason of human beings themselves) may also bear this name subjectively only if they have been drawn out of the universal sources of

> reason, from which critique, indeed even the rejection of what has been learned, can also arise, i.e., from principles. (*CPR* A836-837/B864-865, underlining added, boldfacing in the original)

The second is the equally stark picture of the essentially immature and cowardly person who refuses to acknowledge the fundamental moral idea behind "enlightenment" or *Aufklärung*, which is to think and act for yourself with resolution and courage:

> Enlightenment is the human being's emergence from his self-inflicted immaturity. Immaturity is the inability to use one's own understanding without the guidance of another. This immaturity is self-inflicted if its cause is not lack of understanding, but lack of resolution and courage to use it without the guidance of another. The motto of enlightenment is therefore: *Sapere aude!* Have the courage to use your own understanding! …. [O]nce the germ on which nature has lavished most care—the human being's inclination and vocation to think freely—has developed within its hard shell, it gradually reacts upon the mentality of the people, who thus gradually become increasingly able to *act freely*. Eventually, it even influences the principles of governments, which find that they can themselves profit by treating the human being, who is *more than a machine*, in a manner appropriate to his dignity. (*WE* 8: 35 and 41-42, underlining added)

And the third is the perhaps even starker picture of the person who hides from himself the self-defining fact of his own radical evil, flowing from egoism, by pretending that moral disvalue or wrongness is nothing but bad historical consequences of human activity, and not the direct result of our transcendental freedom and the perversity of the rational "human, all-too-human" will:

> This dishonesty (*Unredlichkeit*), by which we throw dust in our own eyes and which hinders the establishment in us of an authentic moral disposition (*ächter moralischer Gesinnung*), then extends itself also externally, to falsity or deception of others. And if this dishonesty is not to be called malice, it nonetheless deserves at least the name of unworthiness. It rests on the radical evil of human nature which (inasmuch as it puts out of tune the moral ability to judge what to think of a human being, and renders any imputability uncertain, whether internal or external) constitutes the foul stain of our species—and so long as we do not remove it, hinders the germ of good from developing as it otherwise would. A member of the English Parliament exclaimed in the heat of debate: "Every man has his price, for which he sells himself." If this is true (and everyone can decide for himself), if nowhere is a virtue which no level of temptation can overthrow, if whether the good or evil spirit wins us over only depends on which bids the most and affords the promptest pay-off, then, what the Apostle says might indeed hold true of human beings universally, "There is no distinction here, they are all under sin—there is none righteous (in the spirit of the law), no, not one." (*Rel* 6:38-39, underlining added)

Human practical reason is our vital, relatively or absolutely spontaneous, capacity to exercise the power of choice for the sake of instrumental or non-instrumental principles; and

Selbstzufriedenheit is the subjective experience of partially or completely wholeheartedly realizing this capacity. In these ways, inauthenticity in the specifically Kantian sense, or the moral contrary of *Selbstzufriedenheit*, is just to comport yourself "heartlessly," or as if you were nothing but a natural automaton or machine—as if you were nothing but a *Dalek* in the famously long-running British TV science fiction series *Dr Who*[248]—wholly determined by natural causal laws, and neither alive nor practically free. Or in other words, inauthenticity in the Kantian sense is the self-automating denial of your own capacity for practical freedom:

> [I]f the freedom of our will were nothing else than [an *automaton spirituale* when it is impelled by representations]… it would in essence be no better than the freedom of a turnspit, which when once wound up also carries its motions from itself. (*CPrR* 5: 97, underlining added)

The doctrine of *Selbstzufriedenheit*, in turn, highlights the basic way in which my contemporary Kantian theory of free agency transcends Hume's theory of practical agency. For the contemporary Kantian, unlike Hume, practical reason is *not* "the slave of the passions."[249] But this does not imply that for the contemporary Kantian, practical reason is *not* intrinsically connected to our desires, drives, emotions, and feelings, and thus intrinsically connected to our passions. On the contrary, according to my account of practical agency, practical reason *is* intrinsically connected to our passions, and indeed intrinsically connected to our deepest and most self-expressive passions: the ones whose ends we would be prepared to go to the wall for. These passions are *the vital engines* of pure practical reason, and practical reason is *the non-mechanical, relatively or absolutely spontaneous, first-order conscious and also self-conscious structural guidance-and-control system for these engines*. Via our faculty for practical reason, we consciously and self-consciously recognize the relative or absolute objective intrinsic values of ends; and at the very same time and in the same respect, our desires, drives, emotions, and feelings subjectively and more or less wholeheartedly propel us towards those ends by whatever means it rationally takes to get us there.

So we can justifiably defy and deny the standard construal of the internalism-about-practical-reasons-vs.-externalism-about-practical-reasons opposition, which puts Hume's theory of practical agency, as the supposed paradigm of reasons internalism, in diametric and exhaustive opposition to Kant's theory of free agency—or at least to the mainstream contemporary Kantian theory of free agancy, as per Korsgaard or later Parfit—as the supposed paradigm of reasons externalism. For we can can endorse *a uniquely Kantian kind of Frankfurt-style internalism about practical reasons*, which says that all reasons are both justifying (in all three senses) on the basis of objective intrinsic values or ends, and also motivating on the basis of either lower-order or higher-order desires, some of which are innately generated. According to my contemporary Kantian view of practical agency, then, the Categorical Imperative is both *affectively wholly heartfelt* and also *actively known* by rational human animals, which is to say that it is both emotionally and also practically known by means of our faculty of practical reason, which in turn is the same as the faculty of *desire*. In this sense, my non-Korsgaardian, non-Parfitian, non-intellectualist contemporary Kantian theory of practical reasons is perfectly continuous with Hume's theory of internal reasons: although, to be sure, my theory also recognizes a special class of desire-overriding, strictly universal, a priori, categorically normative, non-instrumental internal practical reasons that Hume's theory does not recognize. So to play that Strong Kantian Non-Intellectualist riff on Pascal again, using square brackets:

> The heart has its own [pure practical] reasons that [theoretical] reason knows nothing about.

According to the Strong Kantian Non-Intellectualist theory of free agency that I am expounding and defending here, the wholehearted self-realization of autonomous willing, or autonomous self-fulfillment, is *principled authenticity*. Every time an agent truly acts for the sake of the moral law, she realizes moral worth, and thereby subjectively experiences an aspect of, or some salient degrees of, an ideally complete life of principled authenticity. But if she also thereby achieves some individual or socially-shared human happiness, then she also realizes a proper part of the sole and complete good. Thus according to Kant's ethics as I am understanding it, and according to my non-intellectualist version of contemporary Kantian ethics, there are *two* fundamental values or highest goods:

> (1) *the Highest or Supreme Good*, namely, *the good will*, that is, principled authenticity, and
> (2) *the Sole and Complete Good*, represented by the moral Idea of God,[250] that is, individual and social deep human happiness that is actively guided and controlled by a good will, which thereby expresses an ideal proportionality between moral virtue on the one hand and morally worthy happiness-as-self-fulfillment, spread out over all of humanity, on the other.

In turn, the relation between the Highest or Supreme Good and the Sole and Complete Good is essentialist and mereological. An occurrently autonomous human person's good will, or principled authenticity, is the activating immanent structure (or "essential form") of the vital stuffing (or "prime matter") that is deep rational human happiness, and the living whole that is jointly constituted by them, propagated over all of rational humanity, is the Sole and Complete Good.

Contemporary Kantian ethicists and theorists of practical agency, as Strong Kantian Non-Intellectualists, can therefore be defenders of *strict deontological, non-egoistic, non-act-consequentialist, Existentialist, eudaemonism* in ethics, even despite its being rather a mouthful to say. And in this regard, as in so many others, contemporary Kantian ethics can capture what is most defensible and true in Aristotelian ethics and Humean ethics alike, without collapsing into either egoism, act consequentialism, or classical eudaemonism.

3.5 Conclusion

If the Strong Kantian Non-Intellectualist theory of free agency that I have been spelling out and defending in this chapter is objectively true, as I think it is, then rational human animals, real persons, really and truly possess the kind of metaphysically robust freedom of the will—deep freedom, ultimate sourcehood, or up-to-me-ness—that fully supports moral responsibility in particular, but also fully supports the capacities for Kantian autonomy and principled authenticity. Then the Sole and Complete Good for real human persons is all of us, singly (via principled authenticity) and collectively (via social anarchist mutual aid and political solidarity[251]), getting what we most deeply want, in a way that is actively guided and controlled

by the Highest or Supreme Good, namely, a good will, according to the Categorical Imperative. This, in turn, is achieved by means of the innate dispositional emotion of respect which, when triggered, spontaneously generates a consciously experienced second-order volition that constitutes our higher-order wholehearted desire to achieve morally principled self-transcendence. So according to the Strong Kantian Non-Intellectualist theory of free agency that I am presenting and defending, the passions "are, and only ought to be," *not the Humean enslavers of our rationality*, but instead *the vital engines of our pure practical reason*.

Chapter 4

NEITHER/NOR: THE NEGATIVE CASE FOR NATURAL LIBERTARIANISM

> [B]ecause in self-consciousness the will is known directly and in itself, there also lies in this consciousness the consciousness of freedom. But the fact is overlooked that the individual, the person, is not will as thing-in-itself, but is *phenomenon* of the will, and is as such determined. It has entered the form of the phenomenon, the principle of sufficient reason. Hence we get the strange fact that everyone considers himself to be *a priori* quite free, even in his individual actions, and imagines he can at any moment enter upon a different way of life, which is equivalent to saying that he can become a different person. But *a posteriori* through experience, he finds to his astonishment that he is not free, but liable to necessity; that notwithstanding all his resolutions and reflections he does not change his conduct, and that from the beginning to the end of his life he must bear the same character that he himself condemns, and, as it were, must play to the end the part he has taken upon himself.[252]

4.0 INTRODUCTION

The very idea of freedom of the will, when taken together with the very idea of the natural or physical world, jointly constitute a problem that is perhaps the deepest and most difficult of all modern metaphysical problems. Pre-theoretically, as Schopenhauer so aptly describes it, on the one hand we strongly believe ourselves to be free and non-determined. But also on the other hand, taking mechanistic natural science seriously—and in particular, taking deterministic versions of contemporary physics seriously—we also strongly believe ourselves to be naturally determined, and unfree. And, to put it mildly, we cannot easily reconcile these two directly contrary doxic attitudes, pre-theoretical and natural-scientific. Now of course there are also indeterministic versions of contemporary physics, and, in view of quantum mechanics, especially including Heisenberg's Uncertainty Principle and Bohr's Complementarity Principle, at least micro-level indeterminism seems factually true. But as we have already seen, the thought that we might be *indeterministic* automata is as apt to violate our fundamental concept of ourselves as deeply free, as the thought that we are deterministic automata. So to put this updated Schopenhauerian worry in the Sellarsian terms I used in the Introduction: According to The Manifest Image, we strongly believe ourselves be deeply free and not naturally mechanized; yet according to The Scientific Image, we also strongly believe ourselves

to be really unfree and naturally mechanized, and therefore believe ourselves to be deterministic or indeterministic biochemical puppets and moist robots. But seemingly, it is impossible to fuse the two Images into one. That is the problem of free will in a wordbite.

In this chapter, now re-framing the fundamental metaphysical issues about free agency in terms of the contemporary *non*-Kantian debate about it, as opposed to framing it in specifically *Kantian* terms, as I did in the last two chapters, I will re-describe the nature and implications of the problem of free will as compactly, clearly, and distinctly as I can. One main conclusion I shall draw is that the problem of free will is intimately and indeed inseparably intertwined with at least three other fundamental problems of metaphysics. So "the" problem of free will is really *four* problems about free will. That's the bad news.

The good news, however, is that when we seriously think about the problem of free will as a single set of four inseparably intertwined free will problems—which I call *The Fourfold Knot of Free Agency*—and not just as one independent problem apart from the other three problems, then I think that at least the outlines of an adequate, complete solution to all four of the inseparably intertwined free will problems will emerge. If I am correct about this, then, ironically, one of the biggest problems with the classical problem of free will was our conceptual isolation of it from the other fundamental problems, in a well-intentioned philosophical attempt to solve it by an otherwise perfectly reasonable strategy of conceptually-divide-and-explanatorily-conquer.

The other main conclusion I shall draw is that the four free will problems that make up The Fourfold Knot of Free Agency all loosen up quite radically when we take *biology* at least as seriously as we take physics, and we also treat the phenomenon of life according to the anti-mechanistic, non-reductive, non-dualist, *dynamicist* model of life, along with its background theory of non-equilibrium/complex systems thermodynamics, as I proposed in chapter 2. Then, for all creatures inherently capable of wholehearted autonomy or principled authenticity, namely rational minded animals like us, there *cannot be* freedom-in-natural-mechanism, since the far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, inherently non-mechanical, naturally purposive or naturally teleological process of organismic life at the source of agency is a necessary condition of free agency. And without it we would be nothing but biochemical puppets and moist robots, necessarily lacking deep freedom and therefore without free agency. Nevertheless, there *can* be deep freedom, as, precisely, freedom-in-life. That, again, is The Freedom-in-Life Thesis.

Here is the overall plan of this chapter. In section 4.1, I work out a rationally intuitive definition of free will that can be initially and minimally accepted by all contemporary parties to the debate about the problem of free will, in the sense that they can all reasonably agree that *if* free will really exists, *then* it will have these definitional features. In section 4.2, I describe four fundamental metaphysical threats to the very idea that free will in this intuitive sense actually does or even really possibly can exist, namely:

(1) Universal Natural Determinism,
(2) Fatalism,
(3) Universal Natural Indeterminism, and
(4) Natural Mechanism.

And in section 4.3, I describe a fivefold array of classical and/or standard metaphysical responses to these four threats, namely:

(1) Hard Determinism,
(2) Soft Determinism,
(3) Classical Libertarianism (including its agent causal, non-causal, and event-causal indeterminist sub-kinds),
(4) classical Compatibilism, and
(5) classical Incompatibilism.

Then I make a step-by-step negative case for Natural Libertarianism, by showing that we have good reasons for rejecting both classical Compatibilism and classical Incompatibilism, as well as good reasons for rejecting each of what I will call "The Three Standard Options" of Hard Determinism, Soft Determinism, and Classical Libertarianism.

The first step, in section 4.4, is to undertake a critical examination of some important arguments for classical Incompatibilism. The second step, in section 4.5, is to work out three different, although obviously not wholly unrelated, arguments for my own non-classical version of Incompatibilism, namely, what I call *local incompatibilism with respect to natural mechanism*. The third step, in section 4.6, is to reconsider classical Compatibilism, and extract what I take to be its most philosophically plausible features, which then yields my own correspondingly non-classical version of Compatibilism, which I call *non-local compatibilism with respect to natural mechanism*. And the fourth and final step, in section 4.7, is to examine and then criticize Classical Libertarianism and Hard Determinism alike.

After all that critical negativity, real-metaphysical constructive positivity will then make a full reappearance, like a DC Comics superhero. In the next chapter, chapter 5, I will argue that we have good reasons for retaining some critically qualified features of each of the false classical or standard views—namely, classical Compatibilism, classical Incompatibilism, Hard Determinism, Soft Determinism, and Classical Libertarianism—by incorporating all of those features into a distinctively and indeed radically different successor doctrine: Natural Libertarianism. Natural Libertarianism is radically different precisely because of its dynamicist freedom-in-life doctrine, which fully embeds freedom in physical nature and is robustly pro-science, but also fully excludes natural mechanism at the source of rational animal agency, yet still fully heeds the epistemic, metaphysical, aesthetic, and ethical counsels of natural piety. More generally, I will argue that Natural Libertarianism offers the best overall explanation of all the relevant empirical and rationally intuitive a priori philosophical data about free agency.

4.1 THE INTUITIVE[253] DEFINITION OF FREE WILL

But what *is* the very idea of free will? As I see it, that idea has three basic components.

More specifically, in my opinion, it is rationally intuitive that free will, if it really exists, first, is a rational animal's or real person's choosing or doing things, or refraining from so choosing or so doing, without preventative constraints and without inner or outer compulsion (component 1: *negative freedom*), second, together with the ability to choose or do what she wants, or to refrain from so choosing or so doing (component 2: *positive freedom*). Otherwise put, negative freedom is "freedom-from": if you have negative freedom, then nothing is stopping you from choosing or doing what you want, or refraining from so choosing or so doing—this is sometimes called "the freedom of indifference"—and also nothing is either internally or externally forcing you to choose or do anything, or to refrain from so choosing or

so doing. By contrast, positive freedom is "freedom-to": the ability to choose or do things, or refrain from so choosing or so doing, as a direct expression of your own desires. And in my opinion, it is also rationally intuitive, third, that necessarily a rational animal or real person *P* can freely choose or do something *X* if and only if *P* is causally responsible and also deeply morally responsible for *X* (component 3: *causal responsibility and deep moral responsibility*).

By "deep moral responsibility" for any choice or action *X*, I mean that *X flows from the agent herself, that is, from the real person she self-identically is*, and that the normative value of *X*, especially any moral value of *X* or of some of *X*'s consequences that there might be, also *attaches to the agent herself*. It should be noted, before going on, that, strictly speaking, deep responsibility need not be *moral* responsibility, if the normative value that attaches to the agent herself is *non-moral*.[254] For example, the creator of a beautiful work of art is deeply responsible for the work and its beauty, even if these facts are essentially artistic/aesthetic and non-moral. I will come back to the fundamentally important parallel between artistic creativity on the one hand, and deeply free, deeply responsible agency on the other, in section 4.8 and chapter 5.

In any case, deep moral responsibility should be carefully distinguished from "shallow moral responsibility," by which I mean *second-or-third-person attributions* of responsibility, especially including "reactive attitudes," or judgments, of blame, praise, resentment, punishment, etc., etc., made by other people, for whatever reason. Since second-or-third-person attributions are always only *more-or-less warranted*, and can even be *completely mistaken*, then it is clear that someone can be deeply responsible for *X*, even if she is not shallowly responsible for *X*, and conversely. Deep moral responsibility is a *real-metaphysical* fact, whereas shallow moral responsibility, for all its everyday importance, is only a *social* fact.

It is a striking and indeed passing strange feature of contemporary intellectual life that virtually all recent and contemporary philosophers of free will, agency, and responsibility believe that moral responsibility is essentially shallow.[255] This consensus, in my opinion, is principally due to two factors:

> (i) the exceptional influence of Peter Strawson's essay, "Freedom and Resentment," and
> (ii) the purely sociological fact that compatibilism/soft determinism about free will and/or responsibility are the *default* positions in recent and contemporary professional academic philosophy.

But as a defender of Natural Libertarianism, for all the reasons provided in chapters 1-5, I reject those default positions. Correspondingly, I hold that the attribution-theoretic approaches to moral responsibility are not so much outright false, as *importantly misguided*, since they focus on merely secondary, derivative facts about responsibility, as if they were primary and primitive. On the contrary, *the primary, primitive fact is the metaphysical fact of deep (non-)moral responsibility*, not the secondary, parasitic social fact of second-or-third-person attributions of responsibility.

Now although deep (non-)moral responsibility normally involves causal responsibility, and although conversely causal responsibility normally involves deep (non-)moral responsibility, nevertheless, strictly speaking, causal responsibility and deep (non-)moral responsibility are logically independent of one another, because

> (i) non-human minded animals, children, non-culpably ignorant people, or temporarily insane people cannot be deeply morally responsible for bad free choices, bad free acts, or bad downstream consequences of their free choices and acts, even though they are causally responsible for those choices, acts, and consequences—for example, as in Alfred Hitchcock's 1945 psychoanalytic thriller *Spellbound*, if some innocent child *X* unintentionally kills another innocent child *Y*, by intentionally pushing him down a bannister that, unbeknownst to child *X*, just so happens to have a cruelly sharp spike at the bottom of it, when *X* and *Y* are playing together, then *X* is causally responsible but not deeply morally responsible for *Y*'s death, and
> (ii) a rational animal's or real person's intentional use of causal mechanisms beyond the intentional body movements of that real person can involve deep moral responsibility, and a rational animal's or real person's unintentional triggering of those causal mechanisms by means of her own intentional body movements can lack deep (non-)moral responsibility—for example, as in Kathryn Bigelow's 2008 film *The Hurt Locker*, if some person *X* sets a booby-trap, land mine, or radio-controlled bomb, which is then unintentionally triggered by some other person *Y*'s intentional body movements, and this kills *Y*, then *X* is deeply morally responsible but not causally responsible for *Y*'s death, and *Y* is causally responsible but not deeply morally responsible for her own death, and
> (iii) mutatis mutandis for deep non-moral responsibility—for example, a child might skillfully execute the simple instructions of a great artist, and thus be causally responsible, but not deeply non-morally responsible, for the creation of that beautiful artwork; conversely, that great artist would be deeply non-morally responsible for the creation of that beautiful artwork, but not causally responsible for it.

Here, then, is what I call *The Intuitive Definition of Free Will*:

> Free will, if it really exists, is a rational animal's or real person's choosing or doing things, or refraining from so choosing or so doing, with negative freedom, positive freedom, causal responsibility, and deep (non-)moral responsibility.

In what follows, I will not challenge this three-component definition.

It should be especially noted, however, that The Intuitive Definition, as I have formulated it, stipulatively rules out correctly applying the label "freedom of the will" to non-rational animals, namely, non-person animals, and also to rational animals or real persons who are temporarily incapable of being deeply (non-)morally responsible for their choices and acts. Or in other words, I am stipulatively ruling out correctly applying the label "freedom of the will" to most non-human minded animals, and also to any rational animals', or persons', choices or acts that are due to their non-culpable ignorance, temporary insanity, manipulation by someone else, or to some other overwhelming compulsive force. I think that all things considered, in most cases, this stipulation will not lead to any conceptual, metaphysical, or moral problems.

But at the same time, independently of that stipulation, it is also extremely important to acknowledge that there are some minded animal intentional agents who can, as a constitutive feature of their agency, choose or do things with negative freedom, positive freedom, and causal responsibility—although *never* with deep (non-)moral responsibility, just because they are *not*

rational animals or real persons. This class of genuine intentional agents includes cats, dogs, horses, and many other species of minded non-human animals, as well as some minded human animals. Moreover, on my view, not all non-human animals are non-persons, and not all real persons are autonomous persons in the Kantian sense, namely, *morally* autonomous persons. Otherwise put, some non-human animals are real persons, and some real persons, althought intentional agents with full moral status, are not morally autonomous intentional agents, and therefore are not capable of deep (non-)moral responsibility. For example, it is arguable that normal third-trimester fetuses, infants, and toddlers are all real persons with full moral status. It is also arguable that Great apes (by which, as I mentioned earlier, I mean non-human members of the biological family Hominidae, including bonobos, chimpanzees, gorillas, and orangutans), and perhaps also dolphins, are real persons with the same full moral status.[256] But because they are not morally autonomous persons, neither normal third trimester fetuses, nor infants, nor toddlers, nor other children, nor Great apes, nor dolphins, are deeply (non-)morally responsible for their free actions, even if they *are* causally responsible. Nevertheless, they are still absolutely, intrinsically, nondenumerably, objectively morally valuable creatures, towards whom morally autonomous persons have certain categorical obligations. By an important contrast, however, some other conscious animals—for example, bats, cats, dogs, and horses—arguably are non-person intentional agents, fully capable of what I call *free volition*, yet they do not have absolute, intrinsic, nondenumerable, objective moral value, although they remain subjects of moral value and also proper targets of our moral concern.[257]

To flag the extremely important point about non-person intentional agents, however, and also for terminological convenience in contexts in which this distinction matters for one reason or another, I will say that such non-rational, non-person, minded non-human or human animal intentional agents have *freedom of volition* or *free volition*, although they do not have *freedom of the will*, *free will*, *practical agency*, or *free agency*. But it remains fundamentally true that all non-human or human non-rational minded animals, other things being equal, really do have freedom of volition and therefore really are intentional agents.[258]

Please also note that I have not said that only rational *human* animals, or real *human* persons, can have free will, practical agency, or free agency. I fully concede and fully recognize not only that it is conceivable, and therefore logically possible, but also really possible, that there are non-human alien rational animals, alien real persons—like, for example, the sympathetic character of Klaatu in the breakthrough 1950s sci-fi classic, *The Day the Earth Stood Still*,[259] or the equally sympathetic character of ET in the eponymous, equally breakthrough 1980s sci-fi classic, *ET, The Extra-Terrestrial*.[260] Nevertheless, in order to keep things relatively simple for the purposes of exposition, I will bracket this point and write mostly as if all the actual and really possibly rational animals or real persons are human. Still, wherever the fact that some real persons are non-human plays a salient role in the discussion, I will also be careful to make my formulations reflect that.

In any case, the stipulative distinction between applications of the label "free volition" on the one hand, and of the labels "free will," "practical agency," and "free agency" on the other, is not at all an unfamiliar or unprecedented conceptual and terminological move in the philosophy of free agency. For example, in book iii of the *Nicomachean Ethics,* Aristotle marks a very similar distinction between

> (i) conscious, desiring non-rational animals, that are capable only of the "voluntary" (*hekousia*), and

(ii) conscious, desiring rational animals, who are also capable of "decision" (*prohairesis*).

So in this respect, as in others, my view of free volition, free will, practical agency, and free agency is significantly neo-Aristotelian in its philosophical orientation, as well as being contemporary Kantian.

Before moving on, it is also crucial to distinguish between

(i) freedom of the will, aka "free will," and
(ii) freedom of action, aka "free action."

Free will fundamentally concerns a rational animal's or real person's capacity for conscious choosing or willing, or refraining from so choosing or so willing. By contrast, free action concerns not only the capacity for choosing but also a rational animal's or real person's ability to move his or her own body, that is, his or her ability to carry out a "basic action"—that is, an action in the doing of which, no other acts are performed[261]—or refraining from so doing. So they can come apart in certain circumstances, and are inherently different. Suppose, for example, that a rational human animal's will is both negatively and also positively free. As Locke pointed out in the Essay concerning Human Understanding, book II, chapter 21, it is possible for an intentional agent's will to be free in both of these senses, even if she is unable to carry out a basic act by moving her own body—for example, if she is paralyzed, tied down, or overwhelmed by some external force. Again, and now appealing to a case of "non-basic action"—that is, an action that is done by means of a basic action involving intentional body-movements—she can freely choose to open a door even if that door is, in fact and unbeknownst to her, locked. So a rational animal or real person can have free will even if she does not have freedom of action, whether basic or non-basic.

On the other hand, however, let us suppose that someone's consciousness, character, affects, desires, emotions, and thoughts have all been necessitated either by The Big Bang (which would be distal determination) or by a more spatiotemporally local deterministic process or state of the physical world (which would be proximal determination), or that these have all been manipulated by an evil super-scientist—as, for example, in The Manchurian Candidate. Or suppose, less fancifully but more tragically, that the relevant rational animal or real person is the temporary or permanent victim of an obsessive-compulsive mental disorder. Since in these ways it is possible for a rational animal or real person to be under a psychological compulsion, even though he is able to move his own body without preventative constraint or external compulsion, it is therefore also possible to have freedom of action without free will.

Freedom of action, as opposed to freedom of the will, in the special context of civil society, where the preventatively constraining or compulsive factors are the choices and acts of others, especially including the egoistic, hedonistic, or otherwise consequentialistic choices and acts of others, is also what Kant calls "external freedom" in The Doctrine of Right in The Metaphysics of Morals (MM 6: 237-238). But the more general point is that mere freedom of action is at best an external relation of the intentional agent to her physical and/or social context and environment. It is then clear that free will is more metaphysically basic than free action, precisely because of its direct necessary connections with practical agency, causal responsibility and deep (non-)moral responsibility, personhood, and rationality, hence with free will's internality to intentional agency. Mere freedom of action or generalized external

freedom, without freedom of the will, would be empty and pointless. It would be free action without free agency, and thus free action without any inherent value or inherent meaning. Otherwise put, a life of free action without free agency would be nothing but the so-called life of a mere biochemical puppet or moist robot: the meaningless, zeroed-out existence of a "hollow man," a "man without qualities."[262]

This metaphysical point indirectly highlights a fundamental flaw in Kant's political philosophy, namely, that according to Kant in the Doctrine of Right, human government, or the State, is fundamentally designed to secure and sustain our mutual freedom of action, or mutual external freedom, by means of its special possession of the power of coercive force. Yet by its very nature, considered on its own, external freedom is meaningless, zeroed-out, hollow, and without qualities. Hence it falls outside the scope of the Categorical Imperative or moral law, and the domain of moral virtue and moral worth. So to the extent that the State and other State-like institutions are specifically designed to secure and sustain, by coercion, what is in-and-of-itself morally worthless, it is directly contrary to the teleology of our moral nature, which is ultimately to exit the "juridico-civil community" in order to belong to a universal "ethical community" (Rel 6: 94-97), and equally to exit our "self-incurred immaturity," in order to achieve a radical, principled, and authentic version of rational individual and social "enlightenment," Aufklärung, to the extent that it is humanly possible (WE 8: 35). So, by passively remaining inside the coercive State and other State-like institutions, and in succumbing to our pathological fear, leavened with Stockholm Syndrome—that is, victims' pathological identification with their oppressors—of Hobbes's "war of all against all" in the pre-state or non-state condition, the "state of nature," then we have in effect sold our priceless practical freedom or autonomy down the river to the State and other State-like institutions for the glittering Mephistophelian promise of "public safety" and external freedom. Indeed, the very idea of the "state of nature," and, directly corresponding to it, the classical Hobbesian cognitive illusion of our inherent egoism and mutual antagonism, both urgently need to be thoroughly philosophically criticized and debunked. But this is another story for another day, that I have worked out in "Exiting the State and Debunking the State of Nature" (THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.1).

4.2 THE FOUR METAPHYSICAL HORSEMEN OF THE APOCALYPSE

The capacity for free will is not a "bonus" capacity, namely, a capacity that some people naturally possess but other people do not naturally possess, like the ability to play a musical instrument well, the ability to do higher mathematics, or the ability to learn foreign languages easily. On the contrary, the capacity for free will is essential to our nature. Indeed, according to The Intuitive Definition, the very idea of free will is so centrally embedded within our self-conception of our own rational animal and real personal lives, that it is metaphysically and morally impossible to conceive of ourselves without it. In the *Groundwork for the Metaphysics of Morals*, Kant captures this indispensable self-conception in the following way:

> Reason must regard itself as the author of its principles independently of all alien influences; consequently, as practical reason or as the will of a rational being it must be regarded of itself as free, that is, the will of such a being cannot be a will of his own

> except under the idea of freedom, and such a will must in a practical respect thus be attributed to every rational being. (*GMM* 4: 448)

Nevertheless, our grasp of this indispensable self-conception is not without some serious cognitive dissonance. This, in turn, is principally because four possible metaphysical scenarios, like The Four Horsemen of the Apocalypse—Conquest, War, Famine, and Death—seem to stand directly and threateningly in the way of our ever being able to give an adequate, complete philosophical or scientific theory that would be able to *explain* and *vindicate with decisive reasons* this indispensable conception of ourselves as possessing free will, as free agents, and as real persons. These metaphysical malefactors are

(1) Universal Natural Determinism,
(2) Fatalism,
(3) Universal Natural Indeterminism, and
(4) Natural Mechanism.

Let us now look more closely at them, one by one.

Universal Natural Determinism is the doctrine that the complete series of settled past events, together with the general causal laws of nature, causally necessitate the existence and specific character of all present and future events, including all the choices and acts of rational animals or real persons. This can be formulated even more carefully. Let us adopt the following symbolic conventions, where 'p' stands for an arbitrarily chosen proposition about the natural world, to the effect that p:

C-NEC: It is causally necessary that
Pa: All settled past events are taken together as a complete series
Ln: All the general causal laws of nature are conjoined
Fp: Every fact that p about every present and future event is fixed

Then Universal Natural Determinism can be explicitly stated as:

(C-NEC) [(Pa & Ln) → Fp]

If Universal Natural Determinism is true, then it specifically follows that whatever I am choosing or doing now is necessitated by The Big Bang—or whatever it was that actually constituted and determined the causal and nomological origins of the physical world, its cosmological expansion, its entropy, and its thermodynamics, in a framework that includes general relativity and quantum mechanics. Furthermore, Universal Natural Determinism entails that:

> Causally necessarily, if any two events E_1 and E_2 have exactly similar pasts, then E_1 and E_2 will also have exactly similar presents and futures.

Let us call this *The Closed Future Rule*. The basic idea expressed by The Closed Future Rule is that the present and future of the larger natural world and all the rational animals or real persons in it, are antecedently fixed with causal necessity, and that natural history and the lives

of persons do not contain any inherently random factors. It also follows directly from Universal Natural Determinism that if some super-duper-scientist—named, say, "Trillian"—were able to know all the relevant natural facts about the past and also all the general causal laws of nature, then she would be able to predict all present and future events a priori with scientific certainty; and in this way, Trillian would be an even smarter super-scientist than Frank Jackson's famous super-scientist Mary,[263] who merely knows all there is to know about the neurophysiology of vision.

For clarity's sake, and also because this is going to be dialectically important in my later discussion, it is crucial to distinguish Universal Natural Determinism from a much stronger doctrine which says that the complete series of settled past events, together with the general causal laws of nature, *logically necessitate* the existence and specific character of all present and future events, including all the choices and acts of rational animals or persons. This is *Fatalism*. Let us also adopt this convention:

L-NEC: It is logically necessary that

Then Fatalism can be explicitly stated as:

(L-NEC) [(Pa & Ln) → Fp]

According to Fatalism, there is no logical contingency whatsoever in the causal processes of natural history, or inside the lives of rational animals or real persons. Otherwise put, according to Fatalism all the *causal* links in nature, or inside us, are also *logically necessary* links. It follows directly from Fatalism that if Trillian were able to know all the relevant settled natural facts about the past, and also all the general causal laws of nature, then she would also be able to predict all present and future events a priori with *logical* certainty.

While Fatalism is consistent with Universal Natural Determinism, nevertheless Universal Natural Determinism does not *entail* Fatalism. You can consistently affirm Universal Natural Determinism and also deny Fatalism. Even if every present and future moment's existence and specific character is in itself logically contingent, in the sense that it logically could have been otherwise, given all the actual settled facts about the past and all the general causal laws of nature, nevertheless Universal Natural Determinism can still be true. Universal Natural Determinism says only that any present or later event in time is causally necessitated to exist and have a certain specific character, given that the past exists in the specific way that it does exist, and given the specific character of the general causal laws of nature. But the past did not logically have to be just that way, nor did the general causal laws of nature logically have to be just that way. To be sure, the logical necessity of the past, and the logical necessity of the general causal laws of nature, are not automatically entailed by Fatalism. Yet they are still consistent with Fatalism.

Moreover, Fatalism does not entail Universal Natural Determinism, on at least one not wholly implausible interpretation of Fatalism. If it turned out that both the past and also the general laws of nature *were* logically necessary—if, in effect, the essence of the physical world directly mirrored a system of classical logic, as for example, in Wittgenstein's *Tractatus Logico-Philosophicus*—then this Ultra-Fatalism could hold true even if Universal Natural Determinism were false. Indeed, in the *Tractatus* Wittgenstein claims that all necessity is logical necessity and that causal necessity is not only impossible but even unintelligible:

> 5.133 All inference takes place a priori.
>
> 5.135 In no way can an inference be made from the existence of one state of affairs to the existence of another entirely different from it.
>
> 5.136 There is no causal nexus which justifies such an inference.
>
> 5.1361 The events of the future *cannot* be inferred from those of the present. Superstition is the belief in the causal nexus.
>
> 6.37 A necessity for one thing to happen because another has happened does not exist. There is only *logical* necessity.[264]

Wittgenstein's extremely interesting philosophical response to his own Ultra-Fatalism is what I will call *Mystical Compatibilism*:

> 6.421 It is clear that ethics cannot be expressed. Ethics is transcendental. (Ethics and aesthetics are one.)
>
> 6.423 Of the will as the subject of the ethics we cannot speak. And the will as a phenomenon is only of interest to psychology.
>
> 6.43 If good or bad willing changes the world, it can only change the limits of the world, not the facts; not the things that can be expressed in language. In brief, the world must thereby become quite another. It must so to speak wax or wane as a whole. The world of the happy is quite another than the world of the unhappy.
>
> 6.44 The intuition (*Anschauung*) of the world sub specie aeterni is its intuition as a limited whole. The feeling of the world as a limited whole is the mystical feeling.[265]

I will have more to say about Mystical Compatibilism below. But in the meantime, highlighting Wittgenstein's Ultra-Fatalism clearly brings out the crucial point that Universal Natural Determinism is about the *causal-nomological* necessity of the present and future, not about the *logical* necessity of the present and future. Similarly, Universal Natural Determinism cannot logically guarantee that any particular moment of time will actually exist. For all that Universal Natural Determinism says, it is logically possible that the world might never have existed. Of course, the world does actually exist now. So either the world always existed, or perhaps the world started to exist and then continued to exist until now, or else the world pops in and out of existence discontinuously. —Or whatever, depending on your favorite cosmology and/or theology. But in any case, it is always logically possible that the world might also fail to exist at any present or later time.

I will mention here in an anticipatory way, in order to return to it when I critically discuss the well-known *Consequence Argument* in the next section, that it is a standard strategy for critics of Universal Natural Determinism, whether intentionally or not, to confuse Universal Natural Determinism with Fatalism, whether "ordinary" Fatalism or Ultra-Fatalism. For example, if someone sincerely says

> "If everything is naturally determined, then whatever has happened, was strictly fated to happen, and whatever will happen, strictly must happen, no matter what I choose or do,"

then he is confusing Universal Natural Determinism with Fatalism, and possibly even with Ultra-Fatalism.

It is equally crucial to distinguish Universal Natural Determinism from yet another stronger doctrine, which says that nature is initially created and also sustained at every later moment by the irresistible causal powers of an all-knowing and all-good deity. This stronger doctrine is *Universal Divine Determinism*, aka "Theological Determinism." While Universal Divine Determinism is both consistent with Universal Natural Determinism and indeed *entails* Universal Natural Determinism as a trivial consequence, nevertheless Universal Natural Determinism does not entail Universal Divine Determinism. Even if an all-powerful, all-knowing, all-good, world-creating, and world-sustaining deity does not exist, Universal Natural Determinism can still be true.

In this connection, and corresponding to the fallacy of confusing Universal Natural Determinism with Fatalism, there is an important two-part fallacy that consists in confusing Universal Natural Determinism with Theological Determinism, and then unsoundly inferring universal moral chaos from the denial of Theological Determinism, which I will dub *Smerdyakov's Fallacy*:

> "If God is dead, then everything is permitted."

Smerdyakov's Fallacy is of course so-dubbed because of this famous passage in Fyodor Dostoevsky's *The Brothers Karamozov*, which I have already cited in section 3.1 above as a stark example of the sort of highly self-deceiving, fallacious normative reasoning that is strongly encouraged by The One-Dimensional Conception of Rational Normativity:

> "Take that money away with you, sir," Smerdyakov said with a sigh.
>
> "Of course, I'll take it! But why are you giving it to me if you committed a murder to get it?" Ivan asked, looking at him with intense surprise.
>
> "I don't want it at all," Smerdyakov said in a shaking voice, with a wave of the hand. "I did have an idea of starting a new life in Moscow, but that was just a dream, sir, and mostly because 'everything is permitted'. This you did teach me, sir, for you talked to me a lot about such things: for if there's no everlasting God, there's no such thing as virtue, and there's no need of it at all. Yes, sir, you were right about that. That's the way I reasoned." [266]

To be perfectly explicit, Smerdyakov's Fallacy consists in

> (i) confusing Universal Natural Determinism with Theological Determinism, and also
>
> (ii) mistakenly assuming the truth of Divine Command Ethics: the doctrine that God creates morality and that whatever God wills to be morally right is morally right just because God wills it.

So Smerdyakov is doubly confused. Moreover, from the standpoint of the existentially-oriented Kantian moral theory I defend in *Kantian Ethics and Human Existence*, especially chapters 2 and 6, the moral significance of someone's sincerely asserting

> "If everything is naturally determined, then whatever has happened, was strictly fated to happen, and whatever will happen, strictly must happen, no matter I choose or do,"

and Smerdyakov's Fallacy are exactly the same. He has thereby given himself a license to choose and do whatever he feels like choosing and doing, or not to so choose or so do, without any regard for non-consequentialist moral principles, and constrained only by natural mechanical general causal laws. He thereby comports himself as if he were nothing but a fleshy deterministic or indeterministic real-world Turing machine, a biochemical puppet or moist robot, running a decision-theoretic program for satisfying egoistic, hedonistic, or otherwise consequentialist desires. As we also saw in section 3.3, this sort of highly self-deceived and highly self-serving reasoning—ironically and tragically enough, only a really and truly free agent could ever engage in this sort of duplicitous thinking—is the quintessence of *inauthenticity* from a Kantian point of view, that is, the self-stultifying, self-automating denial of your own capacity for principled authenticity:

> This dishonesty (*Unredlichkeit*), by which we throw dust in our own eyes and which hinders the establishment in us of an authentic moral disposition (*ächter moralischer Gesinnung*), then extends itself also externally, to falsity or deception of others. And if this dishonesty is not to be called malice, it nonetheless deserves at least the name of unworthiness. It rest on the radical evil of human nature which (inasmuch as it puts out of tune the moral ability to judge what to think of a human being, and renders any imputability uncertain, whether internal or external) constitutes the foul stain of our species—and so long as we do not remove it, hinders the germ of good from developing as it otherwise would. A member of the English Parliament exclaimed in the heat of debate: "Every man has his price, for which he sells himself." If this is true (and everyone can decide for himself), if nowhere is a virtue which no level of temptation can overthrow, if whether the good or evil spirit wins us over only depends on which bids the most and affords the promptest pay-off, then, what the Apostle says might indeed hold true of human beings universally, "There is no distinction here, they are all under sin—there is none righteous (in the spirit of the law), no, not one." (*Rel* 6:38-39)

To keep things as simple as possible, however, in what follows I will generally leave aside the special and subtle metaphysical and moral issues associated with the possibility of either Fatalism or Universal Divine Determinism, and concentrate solely on the doctrine of Universal Natural Determinism whenever I am discussing Determinism.

In any case, by sharp contrast to Universal Natural Determinism, *Universal Natural INdeterminism* is the doctrine that Universal Natural Determinism is false, that all connections between events, including the existence and specific character of the choosings and doings of persons, are the result of chance and governed by general probabilistic or statistical causal laws alone, and that no particular future events can be scientifically predicted with certainty a priori. So even our super-duper-scientist Trillian cannot know the future with certainty, and will

simply have to make educated guesses. In particular, Universal Natural Indeterminism entails that

> Causally necessarily, even if two events E_1 and E_2 have exactly similar pasts, then possibly and with some definite degree of probability, E_1 and E_2 will each have a different present and a different future.

Let us call this *The Open Future Rule*. The basic idea of The Open Future Rule is that the present and future of the physical world, together with all the persons in it, is not antecedently fixed, and that natural history or the lives of rational animals or persons contain some inherently random factors. Assuming the truth of The Open Future Rule, it is metaphysically possible that everything in natural history or the lives of rational animals or real persons is just a series of happenings of more or less random events according to probabilistic or statistical general causal laws. This is Universal Natural Indeterminism. It is crucial to note that Universal Natural Indeterminism is still probabilistically or statistically causally *law-governed*, aka "stochastic," and not in any way the same as *natural pandemonium*, which would be utterly lawless.

In any case, it seems self-evident that if all the choices and acts of rational animals or real persons obey either The Closed Future Rule of Universal Natural Determinism or The Open Future Rule of Universal Natural Indeterminism, then theorists of "deep" or metaphysically robust free will, as opposed to the "shallow," merely psychological free will of "reactive attitudes" and "reasons-sensitive mechanisms,"[267] are in serious trouble. More comprehensively then, what I will again call the doctrine of *Natural Mechanism*[268] holds that

> either (i) Universal Natural Determinism is true,
> or (ii) Universal Natural Indeterminism is true,
> or else (iii) *some* events are inherently deterministically caused as regards their existence and specific character, and the *other* events are inherently indeterministically caused as regards their existence and specific character, and *every* event is *either* inherently deterministically caused as regards its existence and specific character *or* inherently indeterministically caused as regards its existence and specific character.

In other words, Natural Mechanism says that all things in nature, including all rational human animals or real human persons, are nothing but deterministic or indeterministic automata.

In section 2.1 above, I also made the proposal that the underlying logic and mathematics of Natural Mechanism jointly satisfy the Church-Turing Thesis, which identifies effective decidability, recursive functions, and Turing-computability, given the two further plausible assumptions of "causal orderliness" and "decomposability" to the effect that

> (i) the causal powers of any physical realization of an abstract Turing machine are held fixed under our general causal laws of nature, especially including the Conservation Laws, and
> (ii) the "digits" over which the Turing machine computes constitute a complete denumerable set of spatiotemporally discrete physical objects.

More precisely then, as before, I am saying that

> Anything X is a *natural automaton,* or *natural machine*, if and only if
>
> (1) X is constituted by an ordered set of causally-efficacious behaviors, functions, and operations (aka "causal powers"),
> (2) the causal powers of X are necessarily determined by all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, together with all the general deterministic or indeterministic causal laws of nature, especially including the Conservation Laws, and
> (3) X's causal powers are all inherently effectively decidable, recursive, or Turing-computable, given two further plausible assumptions to the effect that
> (3i) the causal powers of any real-world Turing machine are held fixed under our general causal laws of nature, and
> (3ii) the "digits" over which the real-world Turing machine computes constitute a complete set of mathematically denumerable (that is, non-real-number, non-complex-number, non-transfinite) quantities, that is, spatiotemporally discrete, physical objects.

Therefore if Natural Mechanism is true, then all sentient and sapient animals, including ourselves, and all other real persons as well, are, *really and truly*, and perhaps also *nothing but*,[269] fleshy deterministic or indeterministic Turing machines, biochemical puppets, and moist robots.

Roughly sixty years before Turing's breakthrough paper in 1936,[270] in 1874, here is how the ultra-Darwinian biologist Thomas Huxley trenchantly put the very same point:

> The consciousness of brutes would appear to be related to the mechanism of their body simply as a collateral product of its working, and to be completely without any power of modifying that working as the steam-whistle which accompanies the work of a locomotive engine is without influence on its machinery. Their volition, if they have any, is an emotion indicative of physical changes, not a cause of such changes… It is quite true that, to the best of my judgment, the argumentation which applies to brutes holds equally good of men; and, therefore, that all states of consciousness in us, as in them, are immediately caused by molecular changes in the brain substance. It seems to me that in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of the organism. If these positions are well based, it follows that our mental conditions are simply the symbols in consciousness of the changes which take place automatically in the organism; and that, to take an extreme illustration, the feeling that we call volition is not the cause of a voluntary act, but the symbol of that state of the brain which is the immediate cause of that act. We are conscious automata, endowed with free will in the only intelligible sense of that much-abused term—inasmuch as in many respects we are able to do as we like—but nonetheless parts of the great series of causes and effects which, in its unbroken continuity, composes that which is, and has been, and shall be—the sum of existence.[271]

So if Natural Mechanism is true, and if Huxley's physicalist reduction of mental facts to naturally mechanistic fundamental physical facts is *also* true, then all of us rational human animals or real human persons are really *nothing but* fleshy deterministic or indeterministic real-world Turing machines, biochemical puppets, and moist robots, that at best merely epiphenomenally, self-deceptively, and tragically *dream* that we are free agents. But in fact it is metaphysically problematic enough even if it is just really and truly the case that we are fleshy real-world Turing machines, biochemical puppets, and moist robots, *without* physicalist reduction, as I will argue in sections 4.4 and 4.5. For the truth of Natural Mechanism is incompatible with the existence of real free agency.

Relatedly, it is crucially important to recognize that the free agency worry about Natural Mechanism is subtly distinct from, and goes beyond, although it still obviously is importantly related to, and indeed includes, the free agency worry about Universal Natural Determinism. Put very simply, if Universal Natural Determinism is true, then either The Big Bang distally necessitates all my choices and actions, or some other more local environmental physical state of the world or process proximally necessitates them, and thus something else really does all the things I seem to choose and do myself, not me. But if Natural Mechanism is true, then even if the real causal source of my choosings and doings is *indeterministic* and spatiotemporally coincides with me, and even if I thereby have the *illusion* that I am the ultimate source of my choosings and doings, nevertheless I am *not* the ultimate source. That is because if Natural Mechanism is true, then in fact the creature that spatiotemporally coincides with me in that context, and does the relevant causing, is really and truly a fleshy *deterministic or indeterministic* real-world Turing machine, a biochemical puppet and moist robot, and *not* me, a rational human animal or real human person. A naturally mechanical causal source, whether deterministic or indeterministic, is *not* a real agentive causal source and *not* a real personal causal source, because it is *not* a real living organism. So it is one thing for something that is merely made out of human flesh, but is really and truly a fleshy deterministic or indeterministic real-world Turing machine, to be an efficacious causal source, as such, and another thing altogether for someone with a real agentive life of her own, a real living organism with capacities for consciousness, intentionality, and caring, whose inner and outer life has meaning, for better or worse, who is a rational "human, all too human" animal or real human person, to be an ultimate efficacious causal source. A causally efficacious real-world Turing machine is categorically *not* the same as a causally efficacious living rational human animal or real human person.

Granting that characterization of Natural Mechanism, and also paying full attention in this context to the point that rational animals or real persons strictly speaking need not all be human, then the fully generalized problem of free agency is this:

> How can rational animals or real persons really and truly choose or do things, or refrain from so choosing or so doing, with negative freedom, positive freedom, and also causal and deep (non-) moral responsibility, in a physical natural world in which Natural Mechanism is prima facie really possible?

4.3 THE THREE STANDARD OPTIONS, NATURAL MECHANISM, AND THE FOURFOLD KNOT OF FREE AGENCY

In the recent and contemporary philosophical debate about the problem of free will, what I call "The Three Standard Options" are the following doctrines—

> *Hard Determinism*: Free will and Universal Natural Determinism are mutually inconsistent,[272] free will is impossible, and Universal Natural Determinism is true.
>
> *Soft Determinism*: Free will and and Universal Natural Determinism are mutually consistent, free will exists, and Universal Natural Determinism is true.
>
> *Classical Libertarianism*: Free will and Universal Natural Determinism are mutually inconsistent, free will exists, and Universal Natural Determinism is impossible,
>
> either (i) because free will exists as an essential property of a special agent-substance, existing outside the natural causal order or inside that order (classical agent-causationism),
> or (ii) because some indeterministic processes exist in nature and free will is among them (event-causal indeterminism),
> or (iii) because some indeterministic processes exist in nature and free will is not among them because it exists over and above natural processes (non-causal indeterminism).

As I mentioned in section 1.1, the thesis that free will and Universal Natural Determinism are mutually logically or metaphysically consistent is classical Compatibilism. So Soft Determinism is a form of classical Compatibilism. Strictly speaking, classical Compatibilism does *not* require asserting either the existence of free will or the truth of Universal Natural Determinism. These are further substantive metaphysical theses that need to be fully explicated and defended. As a matter of fact, however, most contemporary Compatibilists are also Soft Determinists. But it is worth remembering that classical Compatibilism on its own is only a modal metaphysical thesis about mutual consistency, and furthermore that Soft Determinism does not follow just by conceptual-logical entailment from classical Compatibilism.[273]

One way of seeing this important dialectical point is the recognition that Soft Determinism is in fact a version of *deflationary* Libertarianism, since it postulates the actual existence of a certain metaphysically non-robust and indeed merely psychological and/or epistemic kind of free will, whereas classical Compatibilism is not, in and of itself, a version of *any* kind of Libertarianism. Classical Compatibilism is, rather, only the thesis that Universal Natural Determinism and some or another kind of free will, no matter how deflationary or inflationary this conception of free will might be, are mutually consistent.

As I also mentioned in section 1.1, the thesis that, on the contrary, free will and Universal Natural Determinism are mutually inconsistent is classical Incompatibilism. Like classical Compatibilism, classical Incompatibilism on its own is only a modal metaphysical thesis. So Hard Determinism and Classical Libertarianism are both forms of classical Incompatibilism,

yet also involve further substantive metaphysical claims that need to be fully explicated and defended.

There is a contemporary view closely related to Soft Determinism, defended by John Martin Fischer, which says that although free will and Universal Natural Determinism are mutually inconsistent, nevertheless *moral responsibility*[274] and Universal Natural Determinism are mutually consistent, Universal Natural Determinism is true, and moral responsibility exists even if free will is impossible. This is *Semi-Compatibilism*.[275]

Correspondingly, there is a contemporary view closely related to Hard Determinism, defended by Derk Pereboom, which says that free will and Universal Natural Determinism are mutually inconsistent, that free will is conceptually-logically or metaphysically possible but not actual, and that, given the truth of contemporary physics, it follows that one or another of the three mutually exclusive disjuncts of Natural Mechanism is true. This is *Hard Incompatibilism*.[276]

And finally, there is also another contemporary view closely related to both Soft Determinism and Hard Incompatibilism, defended by Manuel Vargas, which says that despite the fact that classical Compatibilism is true, nevertheless our cultural, intellectual, and social history strongly inclines us to believe in classical Incompatibilism, and then we feel and act accordingly, on the basis of a cognitive illusion: so, rationally, we ought to revise our concepts in order to conform with the compatibilistic facts, and then feel and act accordingly. This is *Revisionism*.[277]

Of course there are many other significant recent or contemporary views about free will:[278] I have been trying, and will continue to try, to address all or at least most of these along the way, in the main text or at least in the footnotes. But for the present purposes, the most important point is that the view about free will that I am proposing and defending, Natural Libertarianism, is essentially distinct from each of The Three Standard Options, from Semi-Compatibilism, from Hard Incompatibilism, and from Revisionism too. The essential distinctness of Natural Libertarianism ultimately depends on three factors.

First, as I spelled it out in section 1.1, Natural Libertarianism critically challenges our commitment to the all-too-familiar and seemingly exhaustive dichotomy between classical Compatibilism and classical Incompatibilism, and instead postulates the two specially restricted theses of Non-Local Compatibilism and Local Incompatibilism. As we will remember, Non-Local Compatibilism tells us that some or even most, but not all, of physical nature is made up of deterministic natural automata or machines, whereas Local Incompatibilism tells us that at least some but not all of physical nature is made up of free agents who are themselves nothing more and thing less than human conscious, intentional, caring, rational animals or real persons, that is, far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned thermodynamic systems, which in turn are categorically not deterministic natural automata or machines. According to Natural Libertarianism, very simply put, free agency is just the following two-part fully natural fact:

> (i) not every dynamic system is either a deterministic or indeterministic natural automaton, and
> (ii) at least some of those non-mechanical dynamic systems are living systems, but not merely living systems, and also conscious, intentional, and caring systems, but not merely conscious and caring, intentional systems, because they are also

> rational animals, according to the Two-Dimensional Rational Normativity conception, and, correspondingly, they have real personal lives of their own, created by their deep freedom.

Natural Libertarianism also challenges our commitment to the all-too-familiar and seemingly exhaustive trichotomy between The Three Standard Options of Hard Determinism, Soft Determinism (namely, Deflationary Libertarianism), and Classical Libertarianism (including its classical agent-causal, event-causal indeterminist, and non-causal sub-versions). This is because Natural Libertarianism thereby also indirectly raises its own deeply important, leading, and provocative "revisionist"—really, a philosophical liberationist—question:

> How and why did we ever manage to lock ourselves culturally, intellectually, aesthetically, morally, socially, and politically into this false dichotomy and this false trichotomy; and in what ways would we believe differently, and then feel and act accordingly, if we simply gave them up, liberated ourselves, and became not only "natural pietists," but also "incompatibilistic compatibilists," and thus defenders of Natural Libertarianism?

This philosophical liberationist question, in turn, corresponds directly to D.R. Griffin's late 20th century re-formulation of Schopenhauer's 19th century philosophical clarion-call, that I have quoted twice already, but is so important that it bears repeating at least one more time:

> There is an absolute contradiction between the freedom we all presuppose in practice and the implications of ideas that are widely accepted as established scientific fact. Philosophy has no higher calling than to try to resolve this contradiction at the heart of contemporary culture.[279]

Second, Natural Libertarianism is a non-classical version of Libertarianism that is neither inflationary, by postulating the existence of more than is metaphysically necessary to explain deep freedom, nor deflationary, by postulating the existence of less than is metaphysically necessary to explain deep freedom. As being at once non-inflationary and also non-deflationary about free will, Natural Libertarianism hits the just-right metaphysical mean. This is because

> (i) Natural Libertarianism situates deeply free will fully inside nature, in organismic living processes, which is turn are situated fully inside non-equilibrium thermodynamic systems,
> and also (ii) it postulates its irreducibility to either deterministic or indeterministic natural processes,
> alongside (iii) postulating its locally incompatibilistic causal relevance and causal efficacy,
> together with (iv) postulating its non-local compatibilistic integration with deterministic and indeterministic natural processes,

while at the same time combining all of them in a way that effectively vindicates both the causal responsibility and also the deep (non-)moral responsibility of rational animal agents.

Third, Natural Libertarianism bears a special adversarial metaphysical relationship to Natural Mechanism. What is that relationship?

I have defined Natural Mechanism in three slightly different but still necessarily equivalent ways.

First, in section 2.1, I defined it as the thesis that all biological facts and properties are explanatorily and ontologically necessarily determined by the causal behaviors, functions, and operations of fundamentally physical facts and properties.

Second, also in section 2.1, I defined it as the thesis that every causal behavior, function, or operation in nature, including all the conscious experiences and behaviors of rational animals or real persons, has all its causal powers necessarily determined by the general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, and is Turing-computable from that basis.

And then finally, third, in section 4.2, I defined it as the disjunctive thesis that either Universal Natural Determinism is true, or Universal Natural Indeterminism is true, or else some events are inherently deterministically caused as regards their existence and specific character, and the other events are inherently indeterministically caused as regards their existence and specific character, and every event is either inherently deterministically caused as regards its existence and specific character or inherently indeterministically caused as regards its existence and specific character.

Natural Libertarianism explicitly rejects Natural Mechanism in *all* of its guises.

Now it is quite true, on the one hand, as Kim has pointed out, that an explanatory and ontological appeal to the contemporary natural sciences counts strongly against any version of substance dualism or property dualism.[280] But on the other hand, Natural Libertarianism explicitly rejects both substance dualism and property dualism.

Indeed, even beyond that, defenders of Natural Mechanism cannot rightly claim that their views are inherently more conformable to the contemporary natural sciences than Natural Libertarianism is. This is because Natural Libertarianism is specifically designed to be fully conformable to non-equilibrium thermodynamics, under the non-deterministic interpretation of it offered, for example, by Prigogine,[281] and correspondingly Natural Libertarianism is fully conformable to chemistry, biology, and the cognitive neurosciences, insofar as these are all construed in terms of the non-deterministic interpretation of non-equilibrium thermodynamics and liberal naturalism. In other words, Natural Libertarianism takes natural science seriously too.

More specifically, it is not scientifically unserious to be a liberal naturalist and hold that non-equilibrium thermodynamics, comprehending both physics and chemistry, and biology, especially including organismic biology and ecosystemic biology, and finally cognitive neuroscience, are all anti-mechanistic. Why must all the basic sciences be interpreted in accordance with Natural Mechanism? After all, Church and Turing show us that logical truth in every system at least as rich as classical first-order polyadic quantified predicate logic with identity, aka "elementary logic," cannot be determined by Turing-computable algorithms, and therefore cannot be naturally mechanized; and Gödel's incompleteness theorems show us that every mathematical system at least as rich as Peano arithmetic cannot be naturally mechanized.[282] Yet no one regards elementary logic and Peano arithmetic as somehow less than seriously scientific. If formal piety about logic and mathematics, directly based on the work of Gödel, Church, and Turing, is fully intelligible and defensible, as they surely are, then by the

same token, so too is natural piety about physics, chemistry, biology, and cognitive neuroscience, not to mention the classical "moral sciences."

What defenders of Natural Mechanism can rightly claim at best, is that their view conforms closely to a contemporary physicalist (whether non-reductive or reductive, both of which entail the strong supervenience of everything on the fundamentally physical) conception of nature and the natural sciences. Yet apart from its elective affinity with physicalism, which is a philosophical theory, and not itself a scientific theory, it is hard to know what else could be plausibly said in favor of Natural Mechanism. But in any case, physicalism about logic and mathematics are clearly false.[283] So if one can be fully serious about logic and mathematics without holding either Natural Mechanism or physicalism about them, then by the same token one can fully serious about physics, chemistry, biology, and cognitive neuroscience without holding either Natural Mechanism or physicalism about them, since all of them presuppose logic and mathematics. In particular, if the non-deterministic interpretation of non-equilibrium thermodynamics, together with Church's and Turing's discoveries about logic, together with Gödel's incompleteness theorems, are all true, then Natural Mechanism and physicalism are both false even of physics, and yet we can still be fully serious about logic, mathematics, and physics. Natural Libertarianism clearly meets this theoretical standard.

Back now to Schopenhauer. He evocatively called what he regarded as the three-way fundamental identity between

> (i) the subject of consciousness, intentionality, and cognition,
> (ii) the subject of caring, choosing, and willing, and
> (iii) the subject's phenomenal living body as a proper part of the natural world,

the "world-knot" (Weltknoten).[284] His basic thought was that the phenomenal living human body is nothing more and nothing less than the noumenal Will incarnate:

> As the being-in-itself of our own body, as that which this body is besides being object of perception, namely representation, the *will*, as we have said, proclaims itself first of all in the voluntary movements of this body, in so far as these movements are nothing but the visibility of the individual acts of the will. These movements appear directly and simultaneously with those acts of will; they are one and the same thing with them, and are distinguished from them only by the form of perceptibility into which they have passed.[285]

Leaving aside Schopenhauer's flamboyant subjective-idealist metaphysics, however, the core of the world-knot worry is this: how can the mental subject, the willing subject, and the subject's phenomenal living body as a proper part of the natural world be literally fundamentally *identical* to one another? Correspondingly, and now translating Schopenhauer's worry into contemporary philosophical terms, it seems to me that the problem of free agency is the *deepest* of all metaphysical problems, precisely by virtue of its being inseparably knotted up in a fourfold way with the three other deepest metaphysical problems: the mind-body problem, the problem of mental causation, and the problem of action.

The mind-body problem, as I will understand it here, is this:

> What accounts for the existence and specific character of conscious, intentional, caring minds in a physical natural world?

The problem of mental causation, as I will understand it here, is this:

> What accounts for the causal relevance and causal efficacy of conscious, intentional, caring minds in a physical natural world?

And the problem of action, as I will understand it here, and again paying attention in this context to the point that not necessarily all rational animals or real persons are human, is this:

> What accounts for the categorical difference between the things that rational animals or real persons consciously, intentionally, and caringly do, or perform, and the things that merely happen to them?

Then the fourfold knot of the free agency problem gets twisted up in the following way.

We can start with the mind-body problem. Since the physical world is also the natural world of causally relevant and causally efficacious events in spacetime, then if conscious, intentional, caring minds can be shown to exist and have their specific character in the physical natural world, it follows that conscious, intentional, caring minds must also be causally relevant and causally efficacious in that same physical natural world. That fully invokes the problem of mental causation. Mental causation, in turn, is a necessary condition of intentional action. Mental causation is clearly not sufficient for intentional action, however, because there can be internal psychic compulsions—for example, nervous tics or intrusive voices—that mentally cause behavior and unintentional actions, but do not intentionally cause them. That, in turn, fully invokes the problem of action, since it follows that mental causation is required for the things that rational animals or real persons consciously, intentionally, and caringly do, but does not itself determine the difference between the things that real persons perform and the things that merely happen to them.

Rational animals or real persons have conscious, intentional minds, and above all they care about all sorts of things, other persons, and themselves. So if we cannot explain the existence, specific character, causal relevance, and causal efficacy of conscious, intentional, caring, rational minds in a physical natural world, and if we cannot explain the difference between the things that rational animals or real persons consciously, intentionally, and caringly do, or perform, and the things that merely happen to them, then we certainly also cannot explain how a rational animal's or real person's choosings and doings can ever be negatively or positively free, or include deep moral responsibility. Or more generally, we cannot explain how she can ever be the ultimate source of her choices and acts, so that something is really up to her; and thus we cannot explain how she can ever really be deeply free in a natural world that, at least on the face of it, could be naturally mechanized from top to bottom. At the end of the day, therefore, what we ultimately need to know is the solution to The Fourfold Knot of Free Agency:

> What accounts for the existence, specific character, causal relevance, and causal efficacy of the conscious, intentional, caring minds of rational animals or real persons, insofar as there is a categorical difference between the things that they

> consciously, intentionally, and caringly do, or perform, and the things that merely happen to them, and insofar as they really and truly choose and do things with negative freedom, positive freedom, and also causal responsibility and deep (non-)moral responsibility, in a physical natural world in which Natural Mechanism is prima facie possible?

In turn, it seems to me that the key to an adequate solution to The Fourfold Knot is just this:

> Necessarily, if Natural Libertarianism is true, then Natural Mechanism is false. And Natural Libertarianism is true. So Natural Mechanism is false, biological anti-mechanism is true, and our free agency and our real personhood are the same as our freedom-in-life.

More precisely then, in order to untangle The Fourfold Knot of Free Agency, we need to be able to understand how conscious, intentional, caring, rational animals, or real persons, especially including of course all the human ones, who are deeply free, who are the ultimate sources of their choices and intentional acts, whose choices and acts are thereby up to them, whose choices and acts are more less principled, and more or less wholehearted, are fully-embedded inhabitants of a physical natural world that is not entirely filled with deterministic or indeterministic natural automata, or real-world Turing machines, precisely because it contains some far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned thermodynamic systems—amongst which are rational human animals or real human persons, free agents all.

Here is another crucial point that needs to be emphasized before going on. Let us assume for a moment the truth of Natural Libetarianism. Even assuming that, it is also true that all far-from-equilibrium, asymmetric, complex, self-organizing thermodynamic systems, including free agents like us, can, under some special conditions, accidentally or extrinsically and temporarily approximate, or resemble, the behavior, functions, and operations of deterministic or indeterministic natural automata in those contexts. For example, I can fall off a step-ladder, and be temporarily in the grip of the law of gravity; or I can spin around like a top until I fall down, just like a collapsing puppet with its controlling strings broken; or I can suffer a minor seizure, and be temporarily in the grip of chance or random brain events. In other words, even assuming the truth of Natural Libertarianism, sometimes our activities still approximate to the activities of classical deterministic equilibrium thermodynamic systems; sometimes they still approximate to the activities of thermodynamic systems undergoing "deterministic chaos";[286] and sometimes they still approximate to the throwing of dice. But even fully granting those facts, it does not thereby follow that rational human animals or real persons like us are ever inherently either deterministic or indeterministic natural automata.

This is for two reasons.

First, as I noted in section 1.1, and again in section 2.2, there is, in the nature of things, a fundamental and fully general Kantian distinction between

> (i) an activity's being merely in conformity with (that is, being merely consistent with, acting merely according to) a law or rule, and

> (ii) its being strictly governed by (that is, being strictly entailed or necessitated by, acting strictly from or for the sake of) a law or rule.

And as I also noted in section 2.2, this distinction applies directly and specifically to real-world digital or Turing machine computation, such that there is a basic distinction between

> (i) what is merely correctly describable or can be simulated in Turing-computable terms, and
> (ii) what strictly encodes or inherently implements a real-world-Turing-computable process that flows from all the settled quantity-of-matter-and/or-energy facts about the past, especially The Big Bang, together with the general causal laws of nature, especially including the Conservation Laws.

Therefore, it would be a serious non sequitur to argue from the mere fact that I can be temporarily in the grip of the law of gravity, or temporarily in the grip of "deterministic chaos," or temporarily in the grip of chance or random brain events, to the conclusion that my behaviors, functions, or operations strictly encode, inherently implement, or really incorporate Turing-computable processes in that context or in any other context. Indeed, precisely to the extent that my behaviors, functions, and operations do approximate or resemble Turing-computable processes, then those behaviors, functions, and operations are thereby directly proportionally approaching their being categorically not things that I consciously, intentionally, and caringly do, or perform, but instead they are things that merely happen to me.

Second, as I argued in section 2.6, if we assume the truth of immanent structuralism about properties, my contemporary Kantian theory of mental representation, representational anti-mechanism, and the dynamicist model of life, then it follows that organismic life is both explanatorily and ontologically irreducible to naturally mechanized, Conservation-Law determined, Big-Bang-caused, Turing-computable processes. And since I am a conscious, caring, intentional, rational living organism, it follows that my behaviors, functions, and operations are not inherently naturally mechanized, Conservation-Law determined, Big-Bang-caused, Turing-computable processes.

So I can fall off a step-ladder, and thereby accidentally or extrinsically and temporarily conform to the general causal laws of gravity and falling bodies, break my leg, curse, clutch it, try to stand up, spin around like a top, fall down like a puppet with its controlling strings broken, then faint from the pain, and thereby accidentally or extrinsically and temporarily conform to the activities of classical deterministic equilibrium systems, deterministically chaotic systems, and indeterministic brain events. But those behaviors, functions, and operations are not expressions of my nature as a rational human animal or real human person. In this way, my fall off that step-ladder, my spinning-round like a top and falling down like a broken puppet, and my fainting fit are only constrained or parameterized, and thus necessarily causally enabled, by the complete set of general causal laws governing inherently deterministic and/or inherently indeterministic physical events. Then both the existence and the specific phenomenal characters of my subjective experience of falling off the step-ladder, breaking my leg, cursing, clutching it, etc., and and my own egocentrically-centered point of view on all this, together with all the specific intentional body movements I make along the way, together with the specific non-moral and moral values of my accident, whatever those values may be, are all strictly non-mechanical, non-dualistic facts about my fall, my discomfort, and my fainting fit. They are not

causally entailed or necessitated by all the general deterministic or indeterministic causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, and real-world Turing-computable from that base. For if they had been, then they would not have been my events and my specifics, that is, events and specifics for which I am causally or deeply (non-)morally responsible, by virtue of my consciousness, my intentionality, my caring, my rationality, and my free agency. On the contrary, however, I myself spontaneously bring all those specific phenomenal characters, intentional body movements, and deep (non-)moral values into existence, for better or worse. They are all up-to-me. They are all literally nothing more and nothing less than specific forms of my own real personal life as a free agent, that all grow naturally in my living organismic body.

4.4 THREE ARGUMENTS FOR CLASSICAL INCOMPATIBILISM, AND IN-THE-ZONE COMPATIBILISM

> "Now, my excellent friend," said my companion, "you are in possession of all you need to follow my argument. We see that in the organic world, as thought grows dimmer and weaker, grace emerges more brilliantly and decisively. But just as a section drawn through two lines suddenly reappears on the other side after passing through infinity, or as the image in a concave mirror turns up again right in front of us after dwindling into the distance, so grace itself returns when knowledge has as it were gone through an infinity. Grace appears most purely in that human form which either has no consciousness or an infinite consciousness. That is, in the puppet or in the god."
>
> "Does that mean," I said in some bewilderment, "that we must eat again of the tree of knowledge in order to return to the state of innocence?"
>
> "Of course," he said, "but that's the final chapter in the history of the world."[287]

Now for Step 1 of the negative case for Natural Libertarianism. In the recent and contemporary philosophical literature on free will and moral responsibility, there are two widely-used arguments against classical Compatibilism and in favor of classical Incompatibilism:

(i) The Consequence Argument, and
(ii) The Source Incompatibilist Argument.

I will now look at these two arguments in turn, and then at a third somewhat less well known argument:

(iii) The Causal-Explanatory Exclusion Argument.

The Consequence Argument. The Consequence Argument says this: Assuming that Universal Natural Determinism is true, if we cannot change what is necessarily the case, and we cannot change the past, and we cannot change what is logically entailed by what is necessarily the case, then we cannot change the way we are now or in the future. So we have

no alternative possibilities, and if Universal Natural Determinism is true, then free will is impossible. Therefore, classical Compatibilism is false.

More precisely, however, here is the classical version of The Consequence Argument, as formulated by Peter Van Inwagen.[288] We start by adopting the following conventions, where lower-case 'p' and 'q' stand for arbitrarily chosen propositions to the effect that p and that q about the natural world:

> L-NEC: It is logically necessary that
> Pa: All settled past events are taken together as a complete series
> Ln: All the general causal laws of nature are conjoined
> Fp: Every fact that p about every present and future event is fixed
> (N)p: It is a fact that p and no one has or ever had any choice about whether p

Then we adopt these two inference rules:

> rule α: (L-NEC)p ⊢ (N)p
> rule β: (N)p, (N)(p → q) ⊢ (N)q

Then the overall strategy of the Consequence Argument is two-step. First, it assumes Universal Natural Determinism as a premise, where Universal Natural Determinism is to be understood specifically as

> (L-NEC) [(Pa & Ln) → Fp]

which I will dub

> UND^{PVI}

in order to indicate that it is, specifically, Peter Van Inwagen's formulation of Universal Natural Determinism. And then, second, it derives (N)Fp using only classical propositional logic and the two inference rules α and rule β, as follows:

(1) (L-NEC) [(Pa & Ln) → Fp]	premise, UND^{PVI}
(2) (L-NEC) [(Pa → (Ln → Fp)]	1, propositional logic
(3) (N) [Pa → (Ln → Fp)]	2, rule α
(4) (N)Pa	premise, Unchangeability of the Past
(5) (N) (Ln → Fp)	3, 4 rule β
(6) (N)Ln	premise, Inviolability of the Laws of Nature
(7) (N)Fp	5, 6 rule β

Defenders of classical Compatibilism have replied to The Consequence Argument in various ways.[289] Most of the objections have concentrated on rule β, looking for possible counterexamples to it.[290] In response, defenders of The Consequence Argument have usually replied by refining the characterization of Np and ruling out the counterexamples, thereby re-establishing the soundness of rule β.

Another sort of reply is to challenge (4) or (6), thereby challenging either The Unchangeability of the Past, namely, (N)Pa, or The Inviolability of the Laws of Nature, namely, (N)Ln. The challenge is that there is at least one intelligible and therefore metaphysically possible sense in which we can change the past or violate the laws of nature, consistently with the truth of Universal Natural Determinism, hence classical Compatibilism is defensible.[291] If a "local miracle" happens just before I choose or act, and if causation is the same as "counterfactual influence" (sidebar note: according to the *counterfactual influence theory of causation*, *X* causes *Y* if and only if, necessarily, had some fact about *X* been different from the way it actually was, then *Y* would have been correspondingly different, right up to and including *X*'s non-occurrence determining *Y*'s non-occurrence), then if the miracle had happened differently, I would have chosen or acted differently and the laws of nature would have been different. Correspondingly, there is at least one intelligible and therefore (weakly) metaphysically possible sense in which we can change the past (and the present and the future), since the possible world in which the laws of nature are different will also be a world with a different past (and present and future).

There are two obvious counter-replies to this critical reply to The Consequence Argument.

The first is to provide a new defense of (4), The Unchangeability of the Past. That has been tried, for example, by Wesley Holliday,[292] but still fails. [293]

The second obvious counter-reply is to deny that causation is the same as counterfactual influence. There are three basic sorts of objections here:

> (i) there are "trumping preemption" cases, in which by hypothesis there is counterfactual influence between event *E1* and event *E2*, yet a deviant causal chain constituting an independent event *E3* intervenes so as to be the actual cause of *E2*, thereby "trumping" and "pre-empting" the causal powers of *E1*— for example, if the starter's gun had not gone off, then the sprinter would not have started running, yet the actual cause of the sprinter's starting to run is his being so nervous about the gun going off that he jerks forward just as the gun goes off,
> (ii) causal counterfactuals can hold true in cases that are clearly *not* cases of causation—for example, if 3 + 4 had not equaled 7, then I would not have raised my hand, but 3 + 4's equalling 7 is not clearly the cause of my hand's being raised, and
> (iii) causal counterfactuals fail to hold true in cases of "standard causal overdetermination"—for example, Joe's suspenders cause his trousers to stay up, and Joe's belt causes his trousers to stay up, but it is not the case that if Joe's suspenders had been removed, then his trousers would not have stayed up.

These sorts of worries, in turn, shift the focal point of the debate about The Consequence Argument over to a distinct issue: the acceptability of the counterfactual influence theory of causation.

One possible counter-counter-reply is to these sorts of worries is that *some* sort of counterfactual influence is at least a *necessary* condition of causation, even if it is not also a sufficient condition. For example, in the case of standard causal overdetermination I used just above, it remains true that if Joe's suspenders had been removed and also at the same time Joe's belt had been removed, then Joe's trousers would not have stayed up. And it does seem to generalize that in every genuine case of causation, there is an earlier singular or complex event

such that had it not happened, then the effect also would not have happened. If that is so, then there still remains at least one intelligible and therefore metaphysically possible sense in which we can change the past (and the present and the future) or violate the laws of nature.

OK, *fair enough*, as far as this line of reasoning goes—but that, ultimately, is *not far enough*. I can see some good reasons for thinking that The Consequence Argument is sound, and some other good reasons for thinking that The Consequence Argument is not sound. Nevertheless, it is clear to me, and others, that the recent and contemporary debate about The Consequence Argument is in a vicious dialectical loop, indeed *at stalemate*, with critics and defenders replying, then counter-replying, and then counter-counter replying, pretty much along the lines sketched in the last few paragraphs.[294]

Moreover, in my opinion, the source of this stalemate can be metaphysically diagnosed, and lies in the fact that all standard formulations of The Consequence Argument use the modal operator 'necessarily', aka the box '□', without any careful attention to the different *types* of necessity. But something that is quite evident, when we look back at my initial formulation of Universal Natural Determinism—which, we will remember, is the doctrine that the complete series of settled past events, together with all the general causal laws of nature, causally necessitate the existence and specific character of all present and future events, including all the choosings and doings of persons, or more explicitly formulated,

(C-NEC) [(Pa & Ln) → Fp]

—is that Van Inwagen's classical formulation of The Consequence Argument more or less implicitly construes Universal Natural Determinism as UND^{PVI}, namely, as ordinary *Fatalism*, or (L-NEC) [(Pa & Ln) → Fp].

Now the defender of classical Compatibilism can then easily argue that it is perfectly possible to defend Universal Natural Determinism while rejecting Fatalism. Hence even if The Consequence Argument is sound on some appropriate construal of (N)p for the purposes of defending rule β, nevertheless classical Compatibilism is still possible.

Indeed, this point is also perfectly consistent with the classical Compatibilist argument for the possibility of changing the past or violating the general causal laws of nature. This is because, if Universal Natural Determinism is the doctrine that the complete series of settled past events, together with all the general causal laws of nature, *causally necessitate* the existence and specific character of all present and future events, then since causal necessitation is obviously of narrower modal scope than logical necessitation, it follows that The Unchangeability of the Past and The Inviolability of the Laws of Nature are both implicitly taken to be necessarily true *relative to the general causal laws of nature only*, not *relative to the laws of logic*. From there, it seems an entirely natural step for the defender of classical Compatibilism to interpret the premises (N)Pa and (N)Ln as necessarily true relative to the general causal laws of nature only. After all, the defender of classical Compatibilism might reasonably ask the following rhetorical question:

> "Short of accepting Wittgenstein's Tractarian ultra-Fatalism, how could anyone seriously take either (N)Pa or (N)Ln to be *logically* necessary?"

Then it is obviously logically possible for (N)Pa or (N)Ln to be false, and the classical Compatibilist has won the day.

On the other hand, however, if the defender of The Consequence Argument gets his opponent to agree that Universal Natural Determinism is to be construed as UND^{PVI}, or Fatalism, and then he can also find some reading of (N)p that approximates as closely as possible to implicitly saying, in effect, that

> it is a fact that p and no one has or ever had any choice about whether p, *as a matter of logical necessity*,

without thereby explicitly announcing that he has turned Universal Natural Determinism into Fatalism, then on the contrary the defender of The Consequence Argument will have won the day.

So for these reasons, I think that the stalemate in the debate about The Consequence Argument is permanent, with the defenders of The Consequence Argument always more or less implicitly construing Universal Natural Determinism as Fatalism, while the opponents of The Consequence Argument and defenders of classical Compatibilism always more or less implicitly construing Universal Natural Determinism as a significantly modally weaker thesis than Fatalism. And never the twain shall meet.

The Source Incompatibilist Argument. An importantly different sort of argument strategy, much favored by contemporary post-Frankfurt defenders of classical Compatibilism, is either to concede the soundness of The Consequence Argument, or remain neutral about the stalemated Consequence Argument, and then also argue that both moral responsibility and free will alike do not require alternative possibilities.[295] (I will come back to this crucial issue about alternative possibilities later.) So, according to these contemporary post-Frankfurtians, it is the compatibility of moral responsibility and Universal Natural Determinism that really matters, not the compatibility of free will and Universal Natural Determinism.

Let us suppose for the purposes of argument that these contemporary post-Frankfurt defenders of classical Compatibilism are right that neither moral responsibility nor free will requires alternative possibilities; and let us also suppose for the purposes of argument that the Semi-Compatibilists are right that moral responsibility and Universal Natural Determinism are mutually consistent even if free will does indeed require alternative possibilities and is incompatible with Universal Natural Determinism. Then, on logically independent grounds, The Source Incompatibilist Argument[296] says this: Assuming that Universal Natural Determinism is true, it follows that all my choices and actions are causally necessitated by a series of antecedent events together with the general causal laws of nature. But every such series of antecedent events begins long before I was born, and indeed ultimately begins in The Big Bang. So if Universal Natural Determinism is true, then I am never the ultimate source of my choices or acts, I am never deeply free, and nothing is ever really and truly up to me, hence free will is impossible. Therefore, classical Compatibilism and Semi-Compatibilism are both false.

In my opinion, The Source Incompatibilist Argument is sound. One way of displaying the philosophical power of this argument is as follows. It is rationally intuitive that deep (non-) moral responsibility exists, and also that free will understood as deep freedom, ultimate sourcehood, or up-to-me-ness a priori necessitates deep (non-)moral responsibility. But if Universal Natural Determinism is true, then free will as deep freedom, ultimate sourcehood, or up-to-me-ness is impossible, and nothing else in the world metaphysically suffices to bring deep (non-)moral responsibility into existence. Or in other words, if you take deep (non-)

moral responsibility seriously, then it is very difficult to see how you could consistently be anything but a source incompatibilist.

This remains true even if we fully concede, as I will do later, what contemporary post-Frankfurtians and other "Non-Voluntarists"[297] would have us believe, namely that it is possible to have moral responsibility (whether shallow or deep) in the absence of a free will that requires alternative possibilities. Now Source Incompatibilism, as I am construing it, locates the metaphysical ground of free will in deep freedom, and not in alternative possibilities. But by sharp contrast, the contemporary post-Frankfurtians and other Non-Voluntarists have not provided an account of what constitutes the metaphysical ground of moral responsibility (again, whether shallow or deep) in the absence of a free will that requires alternative possibilities. Rather, they have argued only that (shallow or deep) moral responsibility in this sense is possible and also that, if it exists, then it necessarily involves a suitably-sophisticated moral psychology, complete with the "reactive attitudes" and "reasons-responsive mechanisms."[298] But a sophisticated moral psychology, on its own, does not a metaphysical power-source make. Correspondingly, what Fischer calls "deep control" is at most epistemically deep, and at the same time it is metaphysically shallow, and deflationary, because if Universal Natural Determinism is true, then there are no real minded animal agents existing in the world, and no deep freedom or deep (non-)moral responsibility either.

Seemingly, then, the only prima facie reasonable move left open for the contemporary post-Frankfurt Compatibilist or Semi-Compatibilist to make at this point in the debate, is to propose an "error-theory" or eliminativism about the ordinary concept of moral responsibility.[299] Then this ersatz brand of moral responsibility can still be brought into existence, even in the absence of free will as deep freedom, aka ultimate sourcehood and up-to-me-ness. But, for the contemporary post-Frankfurt Compatibilist or Semi-Compatibilist, that seems wholly self-stultifying. For it is in effect to concede the incompatibility of Universal Natural Determinism and free will as deep freedom, and implicitly accept either Hard Determinism or Hard Incompatibilism.

Of course, it is possible to (try to) be an eliminativist free will skeptic, just as it is possible to (try to) be an eliminativist consciousness skeptic, or an eliminativist moral skeptic.

Elsewhere, I explicitly argue against eliminativist consciousness skepticism and eliminativist moral skepticism.[300] And the very idea of free will, as I spelled it out earlier, presupposes consciousness and morality. So it seems highly unlikely, from that metaphysical vantage point, that eliminativism would hold up for free will. But in any case, eliminativism about free will would also be wholly self-stultifying for post-Frankfurt Compatibilists or Semi-Compatibilists, since they are all also Soft Determinists who think that either free will or moral responsibility, no matter how deflationary and shallow their conceptions of these may be, actually exists, along with Universal Natural Determinism. Indeed, although it is a minimally self-consistent position, it is hard to see what the philosophical point of being a Compatibilist or Semi-Compatibilist and also an eliminativist free will skeptic would truly be.

The Causal-Explanatory Exclusion Argument. But let us imagine for a moment that contemporary post-Frankfurt defenders of classical Compatibilism and Semi-Compatibilism have somehow responded to The Source Incompatibilist Argument in a way that forces at least a tie between Compatibilism or Semi-Compatibilism on the one hand, and classical Incompatibilism on the other. Even so, there is still at least one other powerful argument left in the classical Incompatibilist's repertoire, although this one is somewhat less well-known, because it derives primarily from recent work in the philosophy of mind. This is what is known

as The Causal-Explanatory Exclusion Problem for Mental Causation, developed by Jaegwon Kim.[301]

Kim starts with the assumption that the following principle is undeniably true:

> *The Causal Closure of the Physical*, aka CCP: Only physical things can cause physical things.

Unfortunately, CCP in this version is crucially ambiguous in several respects.[302] So, as I indicated in passing in chapter 1, this is the disambiguated and "precisified" interpretation that I am using:

> CCP: Necessarily, all caused physical events have only event-causes that are consistent with (but *not* necessarily entailed, or otherwise necessitated, by[303]) all the deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or energy facts about the past, especially including The Big Bang.

So understood, CCP rules out any *super*natural or non-spatiotemporal causes, and it also tells us what a thing's *physicality* is:

> Something *X* is physical if and only if *X* has efficacious causal powers that are consistent with (but not necessarily entailed, or otherwise necessitated, by) all the deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or energy facts about the past, especially including The Big Bang.

Granting CCP in this disambiguated and precisified sense, then The Exclusion Problem arises directly from the following principle:

> Since two or more complete and independent causal explanations for the same physical thing cannot exist,[304] there can be only one complete and independent causal explanation of any given physical thing.

This is *The Explanatory Exclusion Principle*. Now "complete" explanations are self-contained and require no other concepts to apply to the relevant physical thing. By contrast, "independent" explanations are complete and also rule out certain other concepts from applying to the relevant physical thing at the same time and in the same respects. To motivate our acceptance of The Explanatory Exclusion Principle, Kim asks us to consider all the possible cases in which there might be two causal explanations respectively invoking *C* (the dualistic mental cause[305]) and *C** (the fundamental physical cause[306]) of the same event *E*:

> (case 1) identity of *C* and *C** (= either reductive materialist type-type identity theory or non-reductive materialist token-token identity theory),
> (case 2) strong supervenience of *C* on *C** (= either reductive functionalism or non-reductive materialism),

> (case 3) *C* and *C** are distinct individually insufficient, individually necessary, and jointly sufficient causes of *E* (= the jointly sufficient mental-and-physical cause theory),
> (case 4) *C* and *C** are different links in the same causal chain leading to *E* (= substance dualist causal interactionism), and
> (case 5) *C* and *C** are distinct individually sufficient causes of *E* (= causal overdeterminationism).

Kim persuasively argues that all the putative cases of dual explanation are either non-independent because the two causal explanations collapse into a single complete and independent causal explanation of *E* (cases 1 and 3), or else they are incomplete because either *C* violates CCP (case 4), or else *C** explanatorily excludes *C* (cases 2 and 5). So The Causal-Explanatory Exclusion Problem is this: Given both CCP and The Explanatory Exclusion Principle, the availability of a physical causal explanation for any physical event effectively excludes any other explanation for that event, and in particular it effectively excludes any *dualistic mental causal explanation* for that event; so efficacious dualistic mental causation of physical events is explanatorily ruled out of court, because the dualistic mental properties of physical events are purely epiphenomenal.

This makes possible a corresponding Causal-Explanatory Exclusion Argument for classical Incompatibilism.[307] That argument says the following. Supposing that Universal Natural Determinism is true, then since by hypothesis each of my choices and actions already has a complete and independent deterministic physical causal explanation, and since both CCP and The Explanatory Exclusion Principle are true, then it follows necessarily that all my dualistic mental choices are purely epiphenomenal and without causal efficacy. So if Universal Natural Determinism is true, and furthermore my free will requires dualistic mental choices, then I am never the sufficient efficacious cause of my choices or actions, hence I am never deeply free or the ultimate source of my choices or acts, nothing is ever up to me, deep (non-)moral responsibility is impossible, and free will is also impossible. Or as Kim aptly puts it:

> Epiphenomenalism strikes most of us as obviously wrong, if not incoherent…. It is the kind of doctrine… that, even if we had to acknowledge it as true, could not serve as a guide to life; it cannot serve as a premise in our practical reasoning, and it is not possible for us to live as though it is true.[308]

In-the-Zone Compatibilism. It is clear that the only possible way a defender of classical Compatibilism can get around both the Source Incompatibilist worry *and also* the epiphenomenalism worry, is to take the philosophical offensive and argue that non-sourcehood and epiphenomenalism are not only perfectly consistent with free will and the ordinary concept of moral responsibility, but also jointly provide us with *exactly the right metaphysics for free will and the ordinary concept of moral responsibility*. In short, you take the bull by the horns, bull-fighter-wise, and flip it over onto its back, so that the seeming big loser in the free will debate is actually the big winner. That could rightly be called *new wave* Compatibilism.

For example, the new wave Compatibilist can argue that the best of all possible rational human lives would be *to be naturally determined to be happy*, that is, to be such that our irreducibly conscious and animal lives fully contain the natural flow of deterministic neurobiological processes towards the very things we both rationally and emotionally need and

want most. So the best of all possible rational human animal or real human personal lives would be to be naturally determined and *in the zone for life*. If this were true, then further rational reflection, deep freedom, deep (non-)moral responsibility, and the autonomous causal efficacy of the consciously mental would only serve *to mess things up and make them worse*, perhaps even radically worse. Taking your life into your own hands in this metaphysically super-robust sense is just too risky, filled with anxiety, uncertainty, and possible suffering, and therefore precisely *not* the "good life" of in-the-zone-for-life happiness, in-the-zone-for-life free will, and in-the-zone-for-life moral responsibility. Therefore, odd as it may at first seem, non-sourcehood and epiphenomenalism are crucial necessary conditions for *just the right kind* of happiness, free will, and the ordinary concept of moral responsibility. Let us call this view *In-the-Zone Compatibilism*,[309] or *In-the-Zone-ism* for short.

I think that In-the-Zone-ism is an extremely interesting view, for all sorts of reasons, and also that it is importantly related to R.E. Hobart's famous argument for the conceptual necessity of Universal Natural Determinism in order to have metaphysically and psychologically robust free will.[310] Even so, I also think that there is something profoundly philosophically mistaken in the very idea of In-the-Zone-ism. This profound mistake can be brought out by recognizing that, looked at historically, In-the-Zone-ism's new wave Compatibilism is *actually* an implicit recurrence to a view originally presented in literary form by the 18th century writer Heinrich von Kleist, in his equally brilliant and strange short story, "On the Marionette Theater."

The historical back-story is that Kleist had been thrown into a serious depression—commonly known as his "Kant-crisis"[311]—by encountering Kant's Critical philosophy, and in particular by encountering Kant's "incompatibilistic compatibilism," according to which we are, at once, phenomenally universally naturally determined and also noumenally free. Kleist fixated on the phenomenal universal natural determinism component, which convinced him that his life was meaningless. In "On the Marionette Theater," he attempted to face up to this meaning-destroying compatibilism by *affirming* the flawless, superhuman character of life as a puppet in an ideal marionette theater, namely, *as a natural automaton*, thereby presenting and defending an early-modern version of In-the-Zone-ism. But his attempt at philosophical self-medication failed, and, tragically, he committed suicide a year later.

The *metaphysical* moral of the tragic Kleist back-story, I think, is that In-the-Zone-ism, along with every other form of classical Compatibilism, can be shown to be false by arguments for what I call *local incompatibilism with respect to Natural Mechanism*. Local incompatibilist, anti-natural-mechanist arguments exploit the profound point I noted in section 4.2 above, which is that it is possible for a creature that has a spatiotemporal coincidence with me, and is even made of human flesh, to be a naturally mechanized causal source of what may *seem* to be my choosings and doings, without its also *being* an ultimate causal source of my choosings and doings. As naturally mechanized, it is nothing but a biochemical puppet and moist robot, a "hollow man," and a "man without qualities." Kleist vividly shows us, however, that this In-the-Zone-ist *conception* can carry an intensely self-medicating aesthetic impact: a life without any risk, a life without the (at times, and especially during a personal crisis) almost unbearable existential anxiety of choice and deep moral responsibility, a life without having to be "human, all too human." The metaphysical point is that causal sourcehood on its own is not sufficient for free will—what is needed for free will is real agentive causal sourcehood and real person-centered causal sourcehood. And that is just as true when a naturally mechanized causal source is "in the zone for life," as when it is not. Indeed, it simply does not really matter to a fleshy deterministic or indeterministic Turing machine whether it is in the zone or not in the zone.

That is because *nothing* ever really matters to a fleshy deterministic or indeterministic Turing machine, one way or the other.

Now moving from 18th century "high" art to 20th century "transgressive" art: Even if it is indeed causally efficacious on its own, a creature that was "in the zone for life" *might just as well be* one of the flesh-eating zombies from George Romero's horror cult classic, *Night of the Living Dead*.[312] The horror we naturally feel, and that Romero's film so brilliantly evokes, when we imagine, or visually represent, a human body that has been galvanized, motorized, and then well-programmed into efficacious movement, from unliving flesh, can be directly aesthetically transferred to In-the-Zone-ism and used to prime our capacity for reliable rational intuition. What, ultimately, leaving aside obvious aesthetic differences, is the deep *metaphysical* difference between Romero's flesh-eating zombie, Kleist's metaphysically-comforting perfect puppet, Harris's biochemical puppet, and Dennett's moist robot? (*None.*) Are we ever even *rationally tempted* to think that a flesh-eating zombie is also a real intentional agent? (*No.*) If not, then why should we ever even be so much as even *rationally tempted* to think that In-the-Zone-ism, or any other kind of classical Compatibilism, no matter how aesthetically self-medicating its conception might be, is a metaphysically adequate doctrine? (*We shouldn't, because even the rational temptation is ultimately rationally self-defeating*).

4.5 Three Arguments for Local Incompatibilism with Respect to Natural Mechanism

Now for Step 2 of the negative case for Natural Libertarianism. In general there is much to say, and in particular there is also much that has already been written by recent and contemporary philosophers, about the opposition between classical Compatibilism and classical Incompatibilism. And, to be sure, many extremely interesting and philosophically rich arguments have been offered on either side of the debate. Two things about this entire debate strike me as quite frustrating and also significantly misleading, however. The first is that it is almost always assumed that free will entails alternative possibilities, that is, the ability to choose or do *X* or *Y*, given all the same facts about the past and the same set of general causal laws of nature. And the second is that it is almost always assumed that the central fact about free will is how it relates to shallow moral responsibility. But both of these common assumptions are false.

Free will does *not* entail alternative possibilities. On the contrary, free will entails only what I call *my capacity for self-commitment to a live option*, or the Kierkegaardian Either/Or, that is, my ability to choose or do something *X*, as opposed to my not choosing or doing *X*, such that *X* would never actually happen (or: have happened) if I were not to choose (or: had I not chosen) *X*, even when, in context, there are no alternative possibilities in the classical sense of branching futures, and the future is temporarily not open.

Moreover, the central or core fact about free will is *not* how it relates to shallow moral responsibility, or even to deep moral responsibility. On the contrary, the core fact about free will is deep freedom, ultimate sourcehood, or up-to-me-ness, which in turn is a metaphysically necessary and sufficient condition of deep (non-)moral responsibility, and yields deep (non-)moral responsibility as a direct and almost trivial consequence, in the context of our inherent capacities for Kantian autonomy and principled authenticity. So in other words, the core fact

about free *will* is free *agency*. I do not mean that deep and specifically moral responsibility is not deeply important. It *is* deeply important—for morality, and morality is of course deeply important for us.[313] And shallow moral responsibility is pretty important too, especially when the police come banging at your door.[314] I mean only that metaphysically speaking, even deep and specifically moral responsibility is a derivative feature of free will that flows from all the primary co-essential features of free agency, including deep freedom, ultimate sourcehood, or up-to-me-ness, together with real personhood (including real personal identity), and the innate capacities for autonomy in the Kantian sense and for principled authenticity. And since shallow moral responsibility is just an epistemic fact that is parasitic on the robust metaphysical fact of deep moral responsibility, then shallow moral responsibility is, at the very least, two degrees of separation away from the core fact about free will, namely free agency.

This point is also clearly indicated by the equally robust metaphysical fact of a kind of responsibility that is also necessarily and sufficiently yielded by free agency, yet is itself *non*-moral: deep *non-moral* responsibility,[315] for example, the responsibility of a creative artist for her artwork. Beyond that, the non-moral deep responsibility of the creative artist *also* provides an analogue or structural metaphor for free agency that is inherently more illuminating than deep *moral* responsibility, the concept of which ties free will, per se, too closely to guidance and evaluation by moral principles. In turn, the concept of the deep non-moral responsibility of the creative artist is infinitely more illuminating than the concept of shallow moral responsibility, which, as attribution-theoretic, misleadingly deflects philosophical attention away from source-incompatibilist egocentrically-centered standpoint of the freely willing first-person agent *herself*, to the displaced-and-eccentric or distanced-and-allocentric epistemic viewpoints of second-persons or third-persons. Correspondingly, my favored meta-philosophical position, which others have aptly called "the primacy of the practical," and which I have also called "the primacy of the normative," in "Kant, Nature, and Humanity" (THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.2) for me means essentially *the primacy of deeply free agency*, not the primacy of morality or moral responsibility, whether deep or shallow.

Deeply free agency, obviously, necessarily includes deep freedom; but somewhat less obviously, the irreducible fact of deep freedom (aka ultimate sourcehood, up-to-me-ness) is jointly constituted by these three elements:

> (i) spontaneous real causality,
> (ii) the capacity for self-commitment to a live option, or the Kierkegaardian Either/Or, which is my capacity to choose or do something X, as opposed to my not choosing or doing X, such that X would never actually happen (or: have happened) if I were not to choose (or: had I not chosen) X, even if, in context, there are no alternative possibilities in the classical sense of branching futures, and the future is temporarily not open, and
> (iii) ownership, which is the fact that my choices and doings belong to me and my life, as a self-identical real person, and do not belong to some other agent or agency.

I will spell out the nitty-gritty details about spontaneous real causality, the live option of self-commitment, and ownership in chapter 5. Nevertheless, for the time being I would like to use this working definition of deep freedom in order to provide three simple arguments for

something I call *local incompatibilism with respect to Natural Mechanism*—that is, for the local incompatibility of deep freedom on the one hand, and Natural Mechanism on the other hand.

By the notion of "local incompatibilism with respect to Natural Mechanism," I mean the mutual inconsistency of deep freedom (aka ultimate sourcehood, up-to-me-ness) and either inherent determinism or inherent indeterminism with respect to the existence and specific character of any of the particular events or processes constituting intentional agency, in the actual event-sequences in which intentional agency occurs. This mutual inconsistency arises as follows. On the one hand, if an agent's particular choices or acts are in fact causally necessitated, with a *closed future*, by all the causal laws of nature, especially including the Conservation Laws, together with the set of settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, in any actual event-sequence in which intentional agency actually or putatively occurs, then she is really a *deterministic* natural automaton in that context, and *deterministically caused* by The Big Bang. Yet if, on the other hand, an agent's particular choices and acts are still causally necessitated by all the general causal laws of nature, especially including the Conservation Laws, together with the set of settled quantity-of-matter-and/or-energy facts about the past, especially The Big Bang, in the actual context in which intentional agency occurs, but instead yield an *open future*, and are inherently the result of chance or random processes under general probabilistic or statistical causal laws, in any actual event-sequence in which intentional agency actually or putatively occurs, then she is really an *indeterministic* natural automaton in that context, and *indeterministically caused* by The Big Bang.

In these ways, natural determinism is the naturally-mechanistic *rock* that binds you hopelessly to the fixed past you can never escape, whereas natural indeterminism is the naturally-mechanistic *hard place* that sacrifices you hopelessly to the random future you can never control. So it's a metaphysical Hobson's Choice: you're damned either way. Either way, even if some creature spatiotemporally coinciding with an actual-sequence real agent is a causal source of what might epiphenomenally *seem* to be real agentially-sourced choosings and doings, but in fact this creature is really a deterministic or indeterministic natural automaton, then that creature is categorically *not* a real agential source in that actual event-sequence, and the creature is *not* deeply free with respect to those particular natural events, because something or someone *other than a real free agent* is ultimately causing these natural events, that is, *The Big Bang is doing it*.

Otherwise put, local incompatibilism with respect to Natural Mechanism is a specially restricted, *actual sequence* version of classical Incompatibilism. And as we saw in section 1.1, and again in section 4.4, Local Incompatibilism is also perfectly coherent with a specially restricted version of classical Compatibilism, namely, Non-Local Compatibilism.

Local incompatibilist arguments, as a species, have their historical origins in an extremely interesting argument that was presented by Kant in the second *Critique*:

> The concept of causality as *natural necessity*, as distinguished from the concept of causality as *freedom*, concerns only the existence of things insofar as it is *determinable in time* and hence as appearances, as opposed to their causality as thing in themselves. Now if one takes the determinations of the existence of things in time for determinations of things in themselves (which is the most usual way of representing them), then the necessity in the causal relation can in no way be united

> with freedom; instead they are opposed to each other as contradictory. For, from the first it follows that every event, and consequently every action that takes place at a moment in time, is necessary under the condition of what was in the preceding time. Now since time past is no longer up-to-me [*in meiner Gewalt*: literally "in my control" or "in my power"], every action that I perform must be necessary by determining grounds *that are not up to me*, that is, I am never free at the point of time in which I act. Indeed, even if I assume that my whole existence is independent form any alien cause (such as God), so that the determining grounds of my causality and even of my whole existence are not outside of me, this would not in the least transform that natural necessity into freedom. For, at every point of time I still stand under the necessity of being determined to an action by *that which is not up to me*, and the series of events infinite a parte priori which I can only continue in accordance with a predetermined order would never begin from itself (*von selbst*): it would never be a continuous natural chain, and therefore my causality would never be freedom. (*CPrR* 5: 94-95, underlining added)

In contemporary philosophy, local incompatibilist arguments have been (at least implicitly, and not under that name) resuscitated by Pereboom and others, and nowadays go under the rubric of "Source Incompatibilism," as I have described it above. Both Kant's and Pereboom's versions of Source Incompatibilism, however, overlook the possibility that Natural Determinism at *some* places and at *some* times and indeed even at *a great many or even most* places and times, is not only perfectly consistent with the truth of local incompatibilism, but in fact is also a necessary enabling condition of deep freedom, as it is conceived by Natural Libertarianism. This, again, is Non-Local Compatibilism.

The other problem with at least *Pereboom's* version of Source Incompatibilism is that it fails to draw the crucial distinction between

> on the one hand, (i) a really possible physical duplicate of a human being like you or like me, materially occupying the same spacetime region, that is operating as a causal source as such, but which can be naturally mechanized, and thus is really a deterministic or indeterministic natural automaton, and
> on the other hand, (ii) a real living, conscious, intentional, caring, rational human minded animal, namely, you or I, who is *not* naturally mechanized, and is therefore *neither* a deterministic *nor* an indeterministic natural automaton, precisely because she is really and truly a far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned thermodynamic system, and a more or less principled, more or less wholehearted, rational human minded animal or real human person, the ultimate agential source of her choices and actions.

Kant, by sharp contrast to Pereboom, seems to have come very close to recognizing this crucial distinction between a really possible counterpart naturally mechanized causal source merely made of human flesh on the one hand, and an inherently agentive living human causal source on the other, in this equally interesting text, also from the second *Critique*, which we have seen already, in part, twice:

> [A]ll necessity of events in time according to natural law can be called the "mechanism of nature," even though it is not to be supposed that things which are subject to it must really be material machines. Here reference is made only to the necessity of the connection of events in a temporal series as they develop according to natural law, whether the subject in which this development occurs be called *automaton materiale* when the machinery is impelled by matter, or, with Leibniz, *automaton spirituale* when it is impelled by representations. And if the freedom of our will were nothing else than the latter, i.e., psychological and comparative and not at the same time transcendental or absolute, it would in essence be no better than the freedom of a turnspit, which when once wound up also carries its motions from itself. (*CPrR* 5: 97, underlining added)

Whether a natural mechanism is a "material automaton" (hence a causal-nomological mechanism) or a "spiritual automaton" (hence a formal mechanism, even if not a causal-nomological mechanism), and even if it either spatiotemporally coincides with me or coincides with me only "spiritually," its putative sourcehood is "no better than the freedom of a turnspit," that is, no better than the pseudo-freedom of a fleshy deterministic or indeterministic real-world Turing machine counterpart version of me, that merely epiphenomenally dreams that it is deeply free and a real person. So Kant clearly sees that all such deflationary conceptions of free will end up by selling selling free agency and real personhood *down the river*, that is, into natural-mechanistic slavery.

Even more importantly for Natural Libertarianism, the crucial difference between Kant and Pereboom here turns on Pereboom's acceptance of a *material-constitution* version of non-reductive physicalism in the philosophy of mind, or what he calls "robust nonreductive materialism."[316] The basic idea behind robust nonreductive materialism is that mental property tokens (events) naturally supervene on physical property tokens (events), and are multiply realizable across distinct physical property tokens. But in sharp contrast to Pereboom's nonreductive materialism, the mind-body relation postulated by Natural Libertarianism is what I call *joint hylomorphic constitution*, which entails the *property fusion* of fundamental biological properties and irreducible, dynamically emergent, immanent structural fundamental mental properties like consciousness, caring, intentionality, rationality, and free will.[317]

An equivalent way of putting the thesis of property-fusion is that conscious, caring, intentional, rational, free minds are nothing more and nothing less than physically irreducible but also non-dualistic, causally efficacious, dynamically emergent, *immanent matter-and/or-energy-structures* of far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned thermodynamic processes. Then, just as the real and complex numbers constructively emerge *within and between* the rational and natural numbers, so too the choices and acts of essentially embodied mindedness and free agency dynamically emerge *within and between* bits of matter and flows of energy inside animals like us, insofar as they are in conformity with all deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, and the Turing-computable algorithms grounded on these. Hence the simpler Conservation-constrained, Big-Bang-primed, Turing-computable structures of matter and/or energy in non-living thermodynamic systems ontologically *give way* in context and over time and *unfold into* the inherently richer dominating non-equilibrium thermodynamic structures of

organismic life, consciousness, intentionality, caring, rationality, and free agency, just as the simpler structures of the rational and natural number systems ontologically *give way* under the mathematical pressures of the power-set operation or the Dedekind-cut method and *unfold into* the inherently richer mathematical structures of the real numbers. As I mentioned before, "giving way and unfolding into" is of course only an illustrative metaphor for the *actualization* of a pre-existing *potential* immanent structure, whether in dynamic emergence or in constructive mathematical emergence, yielding the hylomorphic joint constitution of the resultant novel thermodynamic system or novel number system.

Therefore, the background mind-body metaphysics of Natural Libertarianism entails the non-supervenience of fundamental mental properties on fundamental physical properties, as well as the non-supervience of mental property tokens on fundamental physical property tokens, and also the failure of the multiple realizability of mental property types and tokens over distinct physical property types and tokens. What all this contemporary-metaphysics jargon means in the present setting is simply that for Natural Libertarianism, rational human animals or real human persons will be essentially different from any possible natural automaton counterparts that merely materially occupy the same spacetime regions. Correspondingly, the background contemporary Kantian immanent structuralist, neo-Aristotelian hylomorphic, non-equilibrium thermodynamic mind-body metaphysics that drives Natural Libertarianism's local incompatibilism with respect to Natural Mechanism is *radically more robust* than the so-called "robust nonreductive materialist" mind-body metaphysics that drives Pereboom's version of Source Incompatibilism. "Robust" nonreductive materialism is like "robust" decaffeinated coffee.

In any case, I will now explicitly formulate the three arguments for local incompatibilism with respect to Natural Mechanism.[318] Given my third definition of Natural Mechanism, since it includes both Universal Natural Determinism and Universal Natural Indeterminism as weak disjuncts, it immediately follows that if local incompatibilism with respect to Natural Mechanism is true, then it also entails local incompatibilism with respect to Universal Natural Determinism, and with respect to Universal Natural Indeterminism, alike.

> Argument 1: The Consequence Argument for Local Incompatibilism with Respect to Natural Mechanism
>
> (1) Suppose that Natural Mechanism is true. If all the settled quantity-of-energy facts about the past, especially including The Big Bang, aren't up to me, and if all the deterministic or indeterministic general causal laws of nature aren't up to me, especially including the Conservation Laws, and if causal entailment isn't up to me, then the existence and specific character of whatever is deterministically or indeterministically causally entailed by all the settled quantity-of-matter-and/or-energy facts about the past, together with the general causal laws of nature, are never up to me. (Premise, Suppositional Consequences of Natural Mechanism.)
> (2) Now the settled quantity-of-matter-and/or-energy facts about the past aren't up to me, and the general causal laws of nature aren't up to me, and causal entailment isn't up to me. (Premise, Human finitude.)
> (3) So the existence and the specific character of whatever I'm apparently choosing or doing at any time, which are causally entailed by the settled quantity

of matter-and/or-energy facts about the past, together with the general causal laws of nature, are never up to me. (From (1) and (2), MP.)
(4) So I'm not a free agent. (From (3), and the working definitions of deep freedom and practical agency, hence of free agency.)
(5) Therefore, if Natural Mechanism is true, then I'm not a free agent. (From (1)-(4).)

Argument 2: The Source Incompatibilist Argument for Local Incompatibilism with Respect to Natural Mechanism

(1) Suppose that Natural Mechanism is true. If the ultimate causal source of the existence and the specific character of whatever I'm apparently choosing or doing at any time is The Big Bang, together with all the other settled quantity-of-matter-and/or-energy facts about the past, under either inherently deterministic or inherently indeterministic general causal laws, especially including the Conservation Laws, then neither the existence nor the specific character of whatever I'm apparently choosing or doing at that time is up to me. (Premise, Suppositional Consequences of Natural Mechanism.)
(2) The ultimate causal source of the existence and specific character of whatever I'm apparently choosing or doing at any time is time is The Big Bang, together with all the other settled quantity-of-matter-and/or-energy facts about the past, under all the deterministic or indeterministic general causal laws. (Premise, Natural Mechanism.)
(3) So the existence and the specific character of whatever I'm apparently choosing or doing at any time aren't up to me. (From (1) and (2), MP.)
(4) So I'm not a free agent. (From (3), and the working definitions of deep freedom and practical agency, hence of free agency.)
(5) Therefore, if Natural Mechanism is true, then I'm not a free agent. (From (1) to (4).)

Argument 3: The Deep (Non-)Moral Responsibility Argument for Local Incompatibilism with Respect to Natural Mechanism

(1) If Natural Mechanism is true, then the existence and the specific character of whatever I'm apparently choosing or doing at any time aren't up to me. (From argument 1 above, step (3), and argument 2 above, step (3).)
(2) If the existence and the specific character of whatever I'm apparently choosing or doing at any time aren't up to me, then I'm never deeply (non-)morally responsible for anything. (From the working definitions of deep freedom and practical agency, hence of free agency, and of deep (non-)moral responsibility.)
(3) So if Natural Mechanism is true, then I'm never deeply (non-)morally responsible for anything. (From (1) and (2), CP.)
(4) Therefore, if Natural Mechanism is true, then I'm not a free agent. (From (3), and the working definitions of deep freedom and practical agency, hence of free agency, and of deep (non-)moral responsibility.)

I conclude that classical Compatibilism, Semi-Compatibilism, and Revisionism are all false. In fact, as I will argue in chapter 7 below, there is also a *fourth* argument for local

incompatibilism with respect to Natural Mechanism, from the working definitions of deep freedom and practical agency, hence of free agency, and also of deep (non-)moral responsibility, taken together with the concept of *real personal identity* according to the doctrine of what I call "Minded Animalism." But I will wait until I have explicitly unpacked the notions of real personhood, real personal identity, and Minded Animalism before wheeling out that fourth and final local incompatibilist argument.

4.6 SYMPATHY FOR THE DEVIL: COMPATIBILISM RECONSIDERED

Now for Step 3 of the negative case for Natural Libertarianism. There is something more that needs to be said about the philosophical motives for classical Compatibilism, for which I have a certain amount of sympathy—although, obviously, my sympathy is critically tempered by the three arguments I just gave in section 4.5 for local incompatibilism with respect to Natural Mechanism, not to mention the other three arguments against classical Compatibilism that I spelled out in section 4.4. But as they say, better the devil you know than the devil you don't. What I mean is that even though Local Incompatibilism is true, and therefore classical Compatibilism is false, it also seems to me that classical Compatibilism effectively expresses three very important thoughts, all of which also play directly into my further claim that the specially restricted thesis of *Non-Local* Compatibilism is still true.

The first very important thought in classical Compatibilism, especially in its common guise as Soft Determinism, is that in order to face up to the fact of Universal Natural Determinism and the corresponding theoretical threats of Hard Determinism, Hard Incompatibilism, and the "Semi" part of Semi-Compatibilism as forms of classical Incompatibilism, we must hold that our free will, real human personhood, intentional agency, causal responsibility and (deep[319]) (non-)moral responsibility, consciousness, intentionality, caring, desiring, self-consciousness, and rationality, all inseparably belong to a complete conceptual, metaphysical, and moral package. Lose any one of them and then you lose *all* of them, and thereby you also lose *yourself*. This seems to be a particularly important advantage of classical Compatibilism/Soft Determinism over the "Semi" part of Semi-Compatibilism, which quite unintuitively detaches free will from the rest of the complete package that defines us as *us, rational "human, all too human" animals*.

The second very important thought in classical Compatibilism/Soft Determinism, now in order to face up to the theoretical threat of Classical Libertarianism as another form of classical Incompatibilism, especially in its classical Agent Causation version (for details, see section 4.7 below), is that any adequate account of the metaphysics of free agency must also be logically, metaphysically, and epistemically consistent with the deliverances of the *exact* sciences—that is, mathematics, basic physics, basic chemistry, and basic biology, amongst which, given my earlier arguments, I would obviously want to have special attention drawn to *mathematics* and *biology*.[320] More precisely, defenders of classical Compatibilism/Soft Determinism are saying that our free agency must directly engage with the causally efficacious natural world as it is correctly described by our best mathematical physics. Otherwise our free agency is nothing but metaphysical hocus pocus.[321] Indeed, as we have seen, and as Kim has very effectively shown, without this direct engagement, then the causally efficacious natural world as it is described by

our best mathematical physics both causally and explanatorily *excludes* our free agency, by both causally and explanatorily excluding mental causation.[322]

And the third very important thought in classical Compatibilism/Soft Determinism is a fundamental challenge to any version of Classical Libertarianism, whether classically agent-causal, event-causal indeterminist, or non-causal:

> *If* your version of Libertarianism requires, as a necessary condition of free will, that one must be able to choose (or must believe that you can choose) *Y* as an alternative possibility *even though* everything about one's past, one's character, one's personality, one's desires, the configuration of one's will, and one's patterns of practical knowing and reasoning are centered and focused on one's choosing and doing *X*, so that one's choice of *Y* would be utterly out of character, unwanted, and irrational, precisely because it is merely a matter of *sheer accident or luck*, then your theory of free will is incoherent and false.

In other words, if your version of Libertarianism is either explicitly *indeterministic*, or in any case committed to *alternative possibilities*, then your theory of free will is incoherent and false. Significantly, there is a deep formal analogy between this fundamental luck-based challenge to Libertarianism and the famous *Gettier* fundamental challenge to the classical conception of knowledge according to which knowledge is justified true belief:[323]

> *If* your account of knowledge requires, as a necessary condition of one's justified true belief, that it be possible for the connection between one's conscious-evidence-based reason for believing and the truth of one's belief, to be merely a matter of *sheer accident or luck*, then your theory of knowledge is incoherent and false.

So just as the Gettier challenge clearly and distinctly reveals that classical epistemology must include an *anti-luck condition on justified true believing*,[324] or else give up the "knowledge game," so too this luck-based challenge from classical Compatibilism/Soft Determinism clearly and distinctly reveals that the minimal Libertarian rejection of determinism as incompatible with agency—that is, the miminal Libertarian acceptance of *non*-determinism—must *also* include *an anti-luck condition on free choosing and doing*, or else give up the "agency game," due to the equal and opposite incompatibility of agency with indeterminism.[325]

I fully accept all three of these important thoughts in classical Compatibilism/Soft Determinism, and consider them to be central conditions of adequacy on any correct theory of free agency. Correspondingly, there is at least one non-trivial sense in which Natural Libertarianism is closely akin to classical Compatibilism/Soft Determinism. Given Natural Libertarianism, it is true that in any possible world, if all the settled facts about the past were exactly the same as in the actual world, and if all the general causal laws of nature were exactly the same as in the actual world, and if *I, the agent*, in that world, *am exactly the same as I am in the actual world*,[326] then as a matter of causal necessity all my future choices and acts would also be exactly the same as in the actual world. So Natural Libertarianism, in this special "hyper-actualist" respect, has the same modal profile and modal metaphysics as classical Compatibilism/Soft Determinism. Moreover, it seems to me to be very close to what Nietzsche means by "the myth of eternal recurrence."[327]

The "hyper-actualist" or "eternally recurrent" modal scenario offered by this description of Natural Libertarianism, in effect, collapses all the nomologically possible worlds governed by Universal Natural Determinism onto the one and only *actual sequence*. This actual sequence is indexically fixed by my complete, finite, and unique far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned thermodynamic life as a rational human animal or real human person, up to and including that very moment, including my deep freedom, practical agency, and deep moral responsibility—in short, my free agency. But the crucial point is that the "hyper-actualist" or "eternally recurrent" modal scenario makes all the actual facts about my life, including my free agency, *essential* to my life. Otherwise put, in a certain "hyper-actualist" and "eternally recurrent" sense I am "naturally determined." But at the same time, in this special sense, I am naturally determined *only by myself and by my very own agentive form of life*. Nothing but "me, myself, and I" is determining me: no one and nothing *else* is doing it, and certainly *not* The Big Bang. So my being "naturally determined" in this special sense is really nothing more and nothing less than *my natural self-determination*. My complete, finite, and unique far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned thermodynamic actual-sequence life as a rational human animal or real human person and a free agent, up to and including any given moment, for better or for worse, is not accidentally but instead *essentially* my own life.

This, in turn, yields my response to the luck-based challenge to Classical Libertarianism from classical Compatibilism/Soft Determinism: The anti-luck condition on free choosing and doing is satisfied by Natural Libertarianism, precisely because the agent *herself* is essentially involved in all choosing and doing. That is, whenever the agent chooses or does not choose *X*, or does or does not do *X*, then this fact necessarily flows from the actual-sequence agent *herself* and has nothing inherently to do with classical indeterministic "alternative possibilities." This in turn is because, as the well-known "Frankfurt-style counterexamples" (see section 5.2 below) clearly show, deep (non-)moral responsibility—and thus, according to Natural Libertarianism, deep freedom—are both perfectly possible in the absence, in context, of any classical indeterministic alternative possibilities. The metaphysical bottom-line, then, is that the agent, *as agent*, cannot be alienated from *her own* choices and doings. Agency-undermining luck is thereby ruled out.

As long as that remains true, then there is also no danger of my causal activities and my causal powers being absorbed into the fairly massive deterministic backdrop that necessarily supports all my causally efficacious choices and doings, which would of course mean *my death as a free agent*, by virtue of my being metaphysically converted into a deterministic natural automaton. So although the modal profile and modal metaphysics of this "hyper-actualist" or "eternally recurrent" situation are, strictly speaking, consistent with Universal Natural Determinism, in a space of possible worlds that has been collapsed onto the one and only actual sequence; and although this deeply important fact even further justifies my calling my view *Natural* Libertarianism; of course it is not *really* what the defenders of Universal Natural Determinism have had in mind. —Except, perhaps, for those few non-classical, non-standard defenders of Compatibilism who are also defenders of, or fellow travelers with, Existentialism, for example, Kierkegaard, Nietzsche, the Tractarian Wittgenstein, Sartre, and (in some of his metaphysical moods) Frankfurt. Otherwise put, as a single metaphysical equation:

> "Hyper-actualist" or "eternally recurrent" Compatibilism + Existentialism + The Essential Embodiment Theory + deep freedom + deep (non-)moral responsibility + Kantian practical agency = Natural Libertarianism

But classical Compatibilism/Soft Determinism is *not* the same as "hyper-actualist" or "eternally recurrent" Compatibilism, just as, although for opposite reasons, classical Compatibilism/Soft Determinism is *not* the same as Non-Local Compatibilism. My two-part, rationally charitable, semi-concessive, pro-Compatibilist point here is simply that

> (i) Natural Libertarianism shares *some* non-trivial modal metaphysics with some *non*-classical, *non*-standard versions of Compatibilism/Soft Determinism, and
> (ii) Natural Libertarianism also thereby incorporates *some* non-trivial aspects of classical Compatibilism/Soft Determinism.

And this is sufficient sympathy for the devil.

In any case, and now moving on, if classical Compatibilism/Soft Determinism, Semi-Compatibilism, and Revisionism are all false, due to the truth of local incompatibilism with respect to Natural Mechanism, then that still leaves Classical Libertarianism, Hard Determinism, and Hard Incompatibilism to be critically considered.

4.7 GIVE ME LIBERTY OR GIVE ME DEATH?

Finally, here is Step 4 in the negative case for Natural Libertarianism.

Patrick Henry's notoriously pugnacious *political* libertarian dilemma—Give me liberty or give me death!—of course means: *either* one has political liberty in the 18th century American sense, *or else* life is not worth living. This is clearly a false dichotomy. Even in an 18th century context, it is simply false that the American way of life was the freest way of life. Consider, for example, the moral scandal of slavery. In fact, even in that historical context, there were other ways to be (relatively) politically free in a (relatively) morally acceptable way,[328] without the moral scandal of slavery, and *also* have a life worth living. For example, you could have become a United Empire Loyalist and moved to Canada, like my Hanna forebears did.[329]

Now leaving my family history and mock-serious pro-United Empire Loyalism/Canadianism aside, my serious, basic metaphysical point here is that there is a correspondingly sharp and interestingly analogous dilemma posed by Classical *metaphysical* Libertarians in the free will debate: *either* Classical Libertarianism is true ("live free"), *or else* Universal Natural Determinism is true—and then your rational agentive life is not worth living ("or die!"). My critical response to this Classical Libertarian dilemma is essentially of the same form as my semi-facetious "So why not become a United Empire Loyalist and move to Canada?" response to Henry's political libertarian dilemma. The Classical Libertarian dilemma is clearly a *false metaphysical dichotomy*, because it is entirely possible for you to be deeply free and have a rational agentive life worth living, even if, as I believe, Classical Libertarianism and Universal Natural Determinism are both false. For example, you can affirm Natural Libertarianism. Natural Libertarianism supplies a metaphysically robust conception of free will that is neither *inflated* (dualist) nor *deflated* (reductive or non-reductive physicalist).

Defenders of Classical Libertarianism standardly hold

> either (i) that free agents exist as individual substances and uncaused causes that cause things, whether by means of reasons or not, either (ia) from outside any series of events in spacetime (*classical agent causationism*), or (ib) from inside spacetime, but not by means of event-causation, even including events such as the agent's having reasons (*non-classical agent causationism*),[330]
> or (ii) that free agents exist over and above spatiotemporal nature, and their agency is explained by reasons, always including the belief in alternative possibilities, but they do not cause anything, even by indeterministic means (*non-causalism*[331]),
> or (iii) that free agents exist inside space and time and can cause things by means of indeterministic natural processes (*event-causal indeterminism*[332]).

I will now briefly and critically consider each of these in turn.

The obvious problem with classical agent causationism is that it entails either *dualist interactionism* or *non-standard causal overdetermination*.

According to dualist interactionism, the mental and the physical are essentially distinct kinds of substance that also interact causally. But since all real causal relations (arguably) involve necessitation, and since two essentially different substances can only ever be related *contingently*, it follows that causal interaction between two essentially different kinds of substance is either impossible or else entirely metaphysically mysterious. Furthermore, as Kim has persuasively argued, since according to dualist interactionism, distinct mental substances are immaterial and non-spatial, yet real causal relations clearly require the spatiotemporal individuation of causes and effects, then it seems impossible for it (namely, dualist interactionism) to account for the "causal pairing" of real causes and real effects.[333]

Now it is prima facie arguable, and has in fact been argued by some recent or contemporary philosophers, that *systematic* non-standard causal overdetermination, in which an appeal is made to higher-level "supervenience laws" and "ceteris paribus laws" of the special sciences, avoids violating the causal-explanatory exclusion principle, and can legitimately constitute a second complete and independent causal explanation for a given physical or mental event.[334] But on the contrary, however, no matter how sophisticated the theory of systematic non-standard overdetermination is, nevertheless this theory fails to do the required job. Indeed, Kim's explanatory causal exclusion argument, when supplemented by his supervenience causal exclusion argument, clearly rules out systematic non-standard causal overdetermination.

More explicitly, my basic critical reply to the theory of systematic overdetermination is this: In relevantly similar possible worlds in which all the higher-level supervenience laws or ceteris paribus laws and thus all the higher-level mental properties, that are postulated by the systematic overdeterminationist hypothesis, are all "stripped out," by conceiving them to be missing in that world, then *exactly* the same physical event is still brought about by the lower-level causal laws together with fundamental physical facts, in *exactly* the same lower-level way. It is true that the "fat" possible world, with its higher-level laws and higher-level properties, will be quite different in its *informational and extrinsic structural architecture* from a relevantly similar "skinny" world that lacks these extra laws and properties. But the informational and extrinsic structural architecture, in and of itself, *does no causally efficacious work*: the efficacious causal powers of the physical event itself, together with any other efficacious fundamental physical cause(s) it might have, is essentially the same in the "fat" possible world as it is in the "skinny" possible world. Therefore the addition of higher-level laws and higher-level properties *makes no real or causally efficacious difference whatsoever*

to what actually happens in the actual sequence—although to be sure, again, it will make an informational and extrinsic structural difference to the "fat" world, by making it informationally and extrinsically structurally "fatter." But that difference is *epiphenomenal*. So no matter how sophisticated the systematic non-standard overdeterminationist's theoretical account of causal relevance might be, causal relevance is still not the same as causal *efficacy*. What is needed for the causal efficacy of the mental and free agency is *an immanent or intrinsic structural difference* that spontaneously brings new self-organizing, organismic, and finegrainedly normatively attuned patterns of material movement and/or energy-flow into existence.

Therefore, classical agent-causationist Classical Libertarianism, whether dualist-interactionist or systematic non-standard overdeterminationist, is unacceptable. According to Natural Libertarianism, by sharp contrast, free agents are *not* non-spatiotemporal "metaphysically lonely substances" of any sort: they are instead nothing more and nothing less than immanently-structured life-processes of rational animals, dynamically emergent from the non-mechanical, non-equilibrium thermodynamics of energy-flows.

A similar problem afflicts *non-classical* agent-causationist Classical Libertarianism. Although non-classical agent-causal Classical Libertarians explicitly postulate the existence of agent-substances *inside* spacetime, and thereby avoid Kim's causal pairing problem, they still have a serious metaphysically inflationary problem, which is this. If agent causation takes place from *inside* spacetime, but it cannot occur by means of *events*, even including events such as the agent's having reasons, then how *can* it occur? In short, whereas classical agent causation is a transcendent, or *extra*-spatiotemporal, metaphysical mystery, non-classical agent causation is merely an immanent, or *intra*-spatiotemporal, metaphysical mystery, with no real advance in intelligibility or power of philosophical explanation. All appeals to metaphysical mysteries, transcendent or immanent, are philosophically inflationary—mere postulation of what is explanatorily *otiose*.

Still, there is something philosophically important about "the very idea" of agent causation. Bracketting its problematic commitment to agent-*substances*, the very idea of agent causation is the idea of an *un*determined, *person*-originating, non-spatiotemporal or spatiotemporal, spontaneous, intentional causal source. "Agent causation" is certainly conceptually coherent and intelligible, even if metaphysically mysterious, inflationary, and false. But as a coherent and intelligible conceptual possibility, bracketting its substance-based metaphysics, it shows us that the complete series of settled facts about the past, together with all the general causal laws of nature, can *necessarily underdetermine* rational human intentional agency, which entails *non-determinism* with respect to the agent's choices and acts. Now it would have to be a logically independent thesis, of either classical or non-classical agent-causationist Inflationary Libertarianism, that free will satisfies The Open-Future Rule, which thereby also entails indeterminism. In other words, either classical or non-classical agent-causationist Inflationary Libertarianism is an agency-centered denial of Universal Natural Determinism and consistent with either partial natural indeterminism (that is, the existence of some indeterministic facts or processes in nature) or Universal Natural Indeterminism. But it does not itself *entail* indeterminism—unless, of course, satisfaction of The Open-Future Rule is explicitly added as a further thesis. So *non*-deterministic agency does not, in and of itself, entail *in*deterministic agency.

And that is just as well, since, as we saw above, "indeterministic agency" is in effect an oxymoron, like "accidental knowledge." This is because, when metaphysical indeterminism is included in any version of Libertarianism, it yields *agency-undermining luck*, just as the

classical analysis of knowledge as justified true belief leaves itself open, Gettier-wise, to *knowledge-undermining luck*. Therefore both Universal Natural Determinism and either partial indeterminism or Universal Natural Indeterminism alike can be false, at least with respect to the choices and acts of rational human intentional agents, even though non-determinism is true. In still other words, either partial natural indeterminism or Universal Natural Indeterminism is the *contrary* of Universal Natural Determinism, not its *contradictory*. They cannot both be true, but they can both be false.

This two-part point, namely

> (i) that *non*-deterministic agency does not, in and of itself, entail *in*deterministic agency, and
> (ii) that the very idea of "indeterministic agency" is in itself oxymoronically incoherent and false,

is a somewhat subtle one. But it is also an extremely important, detachable dialectical point for my purposes. This is because, as I have argued already in chapters 2 and 3, and as I will argue again later in chapter 5, the naturally self-determining, far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned thermodynamic events that are inherently characteristic of the lives of rational human animals or real human persons, are inherently both non-*deterministic* and also non-*indeterministic*. For the basic events in our own real agential and real human personal lives are all naturally purposive, or naturally teleological, thereby *underdetermined by* all the general causal natural laws, especially the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially The Big Bang, and inherently uncomputable. So all the real agential events in our lives are *not* naturally mechanized events. Correspondingly, the genuine contradictories in this debate are *not* Universal Natural Determinism and Universal Natural Indeterminism, which are merely contrary versions of Natural Mechanism, but instead *Natural Mechanism* and *Anti-Mechanism*.

By contrast to classical or non-classical agent-causationist Classical Libertarianism, the other two standard versions of Classical Libertarianism—non-causalism and event-causal indeterminism—both strictly entail either a partial version of Natural Indeterminism, namely *the belief in indeterministic alternative possibilities*, or else Universal Natural Indeterminism. Nevertheless, there is also at least one very powerful argument against both Non-Causal Indeterminism and Event-Causal Indeterminism alike. It is another Source Incompatibilist argument, but this time to the effect that indeterminism is as apt to be as inconsistent with deep freedom, ultimate sourcehood, or up-to-me-ness as Universal Natural Determinism is, with respect to the existence and specific character of the particular natural events and processes in the actual sequences in which intentional agency occurs. So it makes essentially the same point as local incompatibilism with respect to Natural Mechanism: If an agent's choices or acts were inherently the result of chance or randomness, even under probabilistic or statistical laws, then those choices and acts would really be the movements of a stochastic natural automaton, biochemical puppet, and moist robot, and not the agent's *own* choosings or doings.[335] In short, either believing in or ontically postulating metaphysical indeterminism at the causal source of intentional action is not only the metaphysically inflationary but also inherently agency-undermining.

Again—there is no good reason whatsoever to think that a creature has free will, just because a significant indeterministic simple or complex event happens *inside her brain*, as opposed to its happening *ten miles away or even on the other side of the universe (as per non-locality effects and entanglement in quantum physics)*, or as opposed to its happening *at the level of microphysical particles and forces*, where her brain does not even exist as a single entity, a living vital organ. Otherwise put, locating a significant indeterministic simple or complex event inside the brain of a mature, healthy human organism does not, in and of itself, confer intentional agency on that organism. Indeed and on the contrary, locating such an event inside the brain of a mature, healthy human organism seems massively more likely to *rob* that human being of her agency, as in the case of a seizure. So merely adding the causal-functional equivalent of a seizure to a creature that is expressing the "reactive attitudes" or implementing a "reasons-sensitive" guidance-control mechanism, *does not an agent make*. This is sometimes called "the disappearing agent problem." But in fact it is simply the fundamental problem of *agency-undermining luck a*gain, now in a slightly different rhetorical guise.[336] Therefore Classical Libertarianism, to the extent that it either believes-in or ontically postulates *in*determinism at the agentive source—as opposed to a mere *non*-determinism that can also equally be consistently combined with non-indeterminism, and with natural purposiveness or natural teleology—and whether this indeterminism is non-causal or event-causal, is ultimately incoherent and false.

4.8 Too Hard to Live with: Strawson's Basic Argument, Hard Determinism, and Hard Incompatibilism

That leaves Hard Determinism and Hard Incompatibilism. Bracketing for the moment the six arguments against classical Compatibilism that I offered in sections 4.4 and 4.5 above, and bracketing in advance the fourth argument for local incompatibilism with respect to Natural Mechanism that I will describe in chapter 7 below, and *also* bracketting the worries I have just rehearsed about Classical Libertarianism, there is at least one prima facie powerful argument in direct support of Hard Determinism, and thereby also directly against Classical Libertarianism. This is what Galen Strawson calls *The Basic Argument*.[337]

The Basic Argument says that in order to be really causally or (deeply) (non-)morally responsible for any event e_1 occurring after time t_1, hence ocurring at time t_2, in addition to e_1 (at t_2)'s being the causal result of the way you are—that is, your nature, character, reasons, and desires—at t_1, you also have to be really and truly causally or (deeply) (non-)morally responsible for a second, prior event e_2 that causes the way you are at t_1, which then leads to a vicious regress, and thus causal responsibility and (deep[338]) (non-)moral responsibility are both impossible. If causal and (deep) (non-)moral responsibility are both impossible, then free will is impossible, Classical Libertarianism is false—indeed any sort of Libertarianism, whether Classical or non-Classical (for example, Natural Libertarianism), is false—Soft Determinism and In-the-Zone Compatibilism are also both false, and even Hard Incompatibilism is false, to the extent that it holds free will to be possible (although not actual). And Hard Determinism is true.

Nevertheless, Strawson's Basic Argument is unsound, and the error occurs in the second step of the argument. It is simply *not* true that in order to be really causally responsible or

(deeply) (non-)morally responsible for any event e_1 occurring after t_1, hence occurring at t_2, then you have to be really causally or (deeply) (non-)morally responsible for an event e_2 that causes the way you are at t_1. To prove this, consider the following analogy between causal or deep *moral* responsibility on the one hand, and artistic creation, together with its deep *non-moral* responsibility, on the other. Everyone will agree that Van Gogh really and truly created some works of art. Indeed, that is precisely what makes those works "authentic Van Goghs." Now suppose that Van Gogh creates a work of art, an "authentic Van Gogh," at t_1. What Van Gogh creates is the result of the way he is *just then*, in the act of creation. So the way he is at t_1 creatively brings about some work of art, an "authentic Van Gogh," at t_2. Then one might assert that in order to be really and truly a creative artist, Van Gogh must also create the way he is at t_1. But that is clearly false. Van Gogh can be (and was) really and truly a creative artist without *creating himself*. Otherwise every creative artist—say, to change examples, Mary Shelley—would have to play the causal role of her own parents, William Godwin and Mary Wollstonecraft, together with the causal role of her own personal history, by bringing herself into existence and then determining the specific character of her entire life up to that point—which is patently absurd.

So, by analogy with artistic creativity and its deep non-moral responsibility, which clearly do *not* require self-creation, The Basic Argument mistakenly lays down an absurdly strong requirement on causal and (deep) (non-)moral responsibility quite generally. In effect, The Basic Argument requires that every causally and (deeply) (non-)morally responsible intentional agent be radically self-causing, or *causa sui*. Nice work if you can get it! Sartre famously wrote, with some truth, that

> To be man means to reach towards being God. Or if you prefer, man fundamentally is the desire to be God.[339]

But even if in fact we all *want* to be God, it certainly would *not* follow that in order to be causally and either deeply *morally* responsible, or deeply *non-morally* responsible, for that matter, we all have to *be* God. Ultimate sourcehood or up-to-me-ness with respect to our choices and intentional acts is perfectly consistent with building on and creatively using, at the current time, and in the thick of things, right there in the actual sequence of events, whatever has already been given to you by physical nature including biology and neurobiology, personal history, rehearsal and training, and cultural history, whether by causal necessity or by chance and randomness, or indeed by your own earlier free agency in the sense specified by Natural Libertarianism, namely, by your freedom-in-life.

In this regard, Kant's description of the mental power of "artistic genius" in the *Critique of the Power of Judgment* (*CPJ* 5: 307-317) provides a fundamentally more accurate and more illuminating analogy for free agency than his own much more famous, but, sadly, *ghost-story-style*, metaphysically mysterious, description of noumenal agency in the third section of the *Groundwork of the Metaphysics of Morals* (*GMM*)—or, for that matter, than John Martin Fischer's interesting but decidedly game-theoretic and algorithm-driven, hence naturally mechanistic, and to that extent, misguided, *card-playing* analogy.[340] In the third *Critique*, Kant writes:

> If, after these analyses, we look back to the explanation … of what is called genius, we find: first, that it is a talent for art, not for science, in which rules that are

> distinctly cognized must come first and determine the procedure in it; second, that, as a talent for art, it presupposes a determinate concept of the product, as an end,, hence understanding, but also a representation (even if indeterminate) of the material, i.e., of the intuition, for the presentation of this concept, hence a relation of the imagination to the understanding; third, that it displays itself not so much in the execution of the proposed end in the presentation of a determinate concept as in the exposition or expression of aesthetic ideas, which contain rich material for that aim, hence the imagination, in its freedom from all guidance by rules, is nevertheless represented as purposive for the presentation of the given concept; finally, fourth, that the unsought an unintentional subjective purposiveness in the free correspondence of the imagination to the lawfulness of the understanding presupposes a proportion and disposition of this faculty that cannot be produced by any following of rules, whether of science or of mechanical imitation, but that only the nature of the subject can produce. According to these presuppositions, genius is the exemplary originality of the natural endowment of a subject for the free use of his cognitive faculties. (CPJ 5: 317-318, underlining added)

As Friedrich Schiller clearly recognized in his Kant-inspired *Aesthetic Letters on the Education of Man*, free agents are *much* more akin to Kantian creative artists than they are to Kantian noumenal angels. Or, I would add, than they are to Fischerian poker-players.

Moreover, and now leaving Kant, Schiller, and Fischer aside, the phenomenon of artistic creativity clearly and distinctly shows us, by analogy, how deep freedom and deep (non-)moral responsibility with respect to choices and acts both strictly require a relatively massive deterministic metaphysical background, in two basic ways.

First, both artistic creativity on the one hand, and also deep freedom/deep (non-)moral responsibility on the other hand, metaphysically require a relatively massive deterministic causal framework that can be exploited by the rational animal or real person in order to create, choose, and act with real causal efficacy. Or in other words, artistic creation and free agency alike both metaphysically need real causal *friction*, real causal *materials*, and real causal *tools* in good naturally-mechanical working order, in order to be able to choose and do things with real causal efficacy.

Second, both artistic creativity on the one hand, and also deep freedom/deep (non-)moral responsibility on the other hand, metaphysically require a relatively massive deterministic backdrop that is definitely *not* itself either the same as, or the product of, rational animals or real persons, in order to be able to *individuate* free agents both epistemically (as believers) and also metaphysically (as real persons), thereby distinguishing them adequately from their larger natural histories and environments, while at the same time still embedding them adequately within their larger natural histories and environments.[341] Otherwise, there is always an epistemic and metaphysical threat that the rational animal or real person will

> either (i) simply merge with her natural history and natural environment, and thereby perish as a free agent, which is the threat of agency-death *by absorption into* pre-existing and ongoing deterministic natural mechanisms,
> or else (ii) simply fail to engage with her natural history and natural environment, and thereby perish as a free agent, which is the threat of agency-death *by alienation from* pre-existing and ongoing deterministic natural mechanisms.

In any case, it seems clearly and distinctly true that deep freedom and deep (non-)moral responsibility are neither self-causation nor self-creation. Consequently, and again by analogy with artistic creation, at t_1 I can be really and truly causally or (deeply) (non-)morally responsible for e_1 at t_2 without having to be causally or (deeply) (non-) morally responsible for the way I am at t_1.

According to Natural Libertarianism, causal responsibility and deep (non-)moral responsibility are produced primarily by *my ability for self-commitment*—just as I am at that moment, or just as I am over that duration of time—to my choice or act. This is a self-commitment that I can also either fail to make or refuse to make, thereby not choosing or doing. So I can either throw myself, just as I am at that moment or over that duration of time, into that choice or act, or not throw myself into it, and the end or goal of that very choice or act would never actually happen without my self-commitment. Therefore there is, in that context, from the agent-centered standpoint, and in the actual sequence of events, at least one thing that is a *live option* for me, and then I can either choose that live option and do it, or not choose or do it. The further fact, if it is a fact, that there are some real alternative possibilities, and thus an open future, is entirely irrelevant from the agent-centered standpoint of deep freedom, although it may well seem to be quite important from a third-person standpoint or either a classical or neo-Hobbesian, neo-Millian individualist (neo)liberal political standpoint.[342]

What ultimately matters metaphysically is only that I can choose or do something—the live option—or not choose or do it, not whether there is anything else I could have chosen or done in that context. Together with the anti-mechanism that is presupposed by it, my capacity for self-commitment to a live option thus satisfies the anti-luck condition on free choosing and doing—a condition that is lacking in Classical Libertarianism—by guaranteeing that all choosings and doings essentially involve the agent herself, necessarily flow from the agent herself, and directly attach to the agent herself, for better or worse.

In this way, free agency is essentially a rational minded animal's *natural creativity*—that is, its self-determining, biologically-grounded, non-equilibrium-thermodynamics-grounded, ability successfully to exploit the "natural open space" available to it in the actual sequence of worldly events and processes, or not, even if, in that context, there are no alternative possibilities and the future is temporarily not open. Indeterminism at the source of agency has nothing to do with it! And that is precisely why Natural Libertarianism is *not* equivalent to Classical Libertarianism, whether in its metaphysical manifestations (agent-causal, non-causal, or event-causal indeterminist) or in its political manifestations (classical liberalism or neoliberalism).

In turn, my capacity for self-commitment to a live option is just the Kierkegaardian Either/Or. Perhaps, as Kierkegaard also says, *I will regret my choice either way*. But that seems excessively cynical—or perhaps it is just appropriately Danish and melancholic. Nevertheless, what does seem to be absolutely true about having the capacity for self-commitment to a live option is that I will thereby be causally and deeply (non-)morally responsible either way. As Kierkegaard puts it:

> On the whole, to choose is an intrinsic and stringent term for the ethical. Wherever in the stricter sense there is a question of an Either/Or, one can always be sure that the ethical has something to do with it. The only absolute Either/Or is the choice between good and evil, but this is also absolutely ethical.[343]

So causal responsibility and deep (non-)moral responsibility are not produced by my causal history—that is, by whatever it is that brings about the way I am at some given moment in time, by bringing about my nature, character, reasons, and desires at that time. Nor are causal responsibility and deep (non-)moral responsibility matters of indeterministic alternative possibilities and branching, open futures. Doubly on the contrary, causal responsibility and deep (non-)moral responsibility are produced solely by my unique appropriation of the way I actually am at some given situational moment in spacetime, or over some duration of time, in the actual sequence of events, and also by my unique personal contribution to, and my more or less passionate involvement in, what I choose and do at that very time, no matter *how* that given situational moment was originally causally brought about. —Provided, of course, that the series of thermodynamic events or processes that causally brought about that very situational moment do not inherently rule out my deep freedom and deep moral responsibility, and also yield appropriate enabling conditions for the efficacious operations of my natural causal powers in that context.

It is crucial to recognize, for the purposes of my later discussion of the capacity for self-commitment to a live option in chapter 5, that exercising the self-commitment capacity need not be in any way occurrently self-conscious, self-reflective, or deliberative. On the contrary, normally and for the most part, exercising the capacity for self-commitment to a live option is *pre-reflectively conscious and spontaneous*. It is the same as what Frankfurt calls conscious "effective first-order desires," insofar as these occur in rational animals or persons. Correspondingly, exercising the capacity for self-commitment to a live option is identical with what Maiese and I, in the theoretical context of action-theory, have called *effortless trying*.[344]

Back now to the final step of the negative case for Natural Libertarianism. Here is another argument against Hard Determinism. If Hard Determinism is true, then real free will is nothing but a myth. There is neither a real fact of free will, nor an intelligible, defensible concept of free will. In other words, Hard Determinism is *radically* deflationary about free will. And in these respects, Hard Determinism plays essentially the same role in the philosophy of free will that *eliminative materialism* does in the philosophy of mind.[345] As I mentioned in passing above, eliminative materialism outright denies and rejects not only the existence of everything mental, including minds, mental states, mental events, mental processes, and mental properties, but also "the very ideas," or concepts, of them. Correspondingly, the prime example of eliminativism is the claim that commonsense or "folk" psychology is nothing but a pseudo-science that must be fully replaced by fundamental physics, chemistry, and biology.[346] So eliminative materialism asserts that because of what the natural sciences tell us, we must learn to live not only without the apparently irreducible *facts* of our own consciousness, intentionality, caring, and rationality, but also without their *concepts*. Similarly, Hard Determinism outright denies the existence of deep freedom, ultimate sourcehood, and up-to-me-ness, and asserts that because of what the natural sciences tell us, we must learn to live without the apparently irreducible facts of our own rationality, real human personhood, intentional agency, and deep (non-)moral responsibility, and also without their concepts.

Well, believe it if you can! On the contrary, however, it seems to me that if that is true, then it is basically The End of the World as We Know It—to borrow an apt turn of phrase from Jerry Fodor and R.E.M.'s lead singer and lyricist, Michael Stipe.[347] What I mean is that I simply cannot see how we could possibly ever rationally get by without the facts or concepts of free agency, deep (non-)moral responsibility, real human personhood, or indeed human rationality itself. Of course, that in and of itself does not prove that Hard Determinism is false. Hard

Determinism might still be true, when it is taken together with a corresponding "debunking strategy" or "error-theory" which purports to show that we are systematically deceived and mistaken about the reality-indicating force of the phenomenology of our own experience of ourselves as rational human free agents—which, in the next chapter, I will call "veridical psychological freedom." It also purports to show that we are systematically deceived and mistaken about the truth of our beliefs in (and, correspondingly, deceived and mistaken about the applicability and possession of our concepts of) free agency, deep (non-)moral responsibility, and real human personhood.

But surely the hypothesis that we have a reality-indicating phenomenology (veridical psychological freedom), and also a true fundamental rational human self-conception, provides a much *better* overall philosophical theory than a debunking strategy or error-theory. This is because the truth of an error-theory requires the fact of a systematic, widespread metaphysical, epistemic, and practical *illusion* about human rationality, which of course in principle equally extends to error-theories and error-theorists too—and that is clearly self-refuting. Since theories are rational human achievements, how could there be real error-*theories* or real error-*theorists* without real human rationality and real rational human animals? So the hypothesis that Hard Determinism and its error-theory are false is *overwhelmingly* more plausible, given everything else that I experience, feel, believe, and know, than the contrary hypothesis that Hard Determinism and its error-theory are correct.

In other words, I think that there is a compelling, G.E. Moore-style argument against Hard Determinism,[348] which provides me with a sufficiently good epistemic reason to believe in, and also a sufficiently good practical reason to work out, a better metaphysical theory about free agency than Hard Determinism. In this sense, Hard Determinism, its radical deflationism, and its error-theory are rationally and humanly just *too hard to live with*; hence I am rationally and humanly entitled to pursue the project of rejecting them and philosophizing otherwise, until such time as my project has been decisively shown to be unintelligible or indefensible.

So much for Hard Determinism. But our critical work in this connection is not *quite* finished. Interestingly and importantly, Pereboom's Hard Incompatibilism does *not* have the entailment that there is no genuine concept of free will.[349] According to the Hard Incompatibilist, real free will is possible, but actually does not exist. So while, according to Pereboom, there is no actual *fact* of real free will, there is nevertheless an intelligible *concept* and also a vivid *phenomenology* of real free will, and correspondingly, these can be taken to support the thesis that, in some non-trivial sense, we really are "deterministic agent-causes."[350] In these respects, Hard Incompatibilism is a significantly more subtle and defensible view than Hard Determinism.

But where Hard Incompatibilism has serious internal difficulties, I think, is in reconciling the explicit rejection of the fact of real free will with an explicit acceptance of the genuine *concept* of real free will, the vivid *phenomenology* of free will, and the thesis about our "deterministic agent-causal" powers. This mixed attitude of rejection/acceptance seems to me to produce serious cognitive (and practical) dissonance. According to Pereboom, I really and truly am a biochemical puppet and moist robot: I know this via mechanistic physics and biology, and "scientific philosophy." Yet at the same time, by virtue of the genuine concept and vivid phenomenology of real free will, I also cannot help believing outside my laboratory or study that I am (in some sense) deeply free and deeply (non-)morally responsible for my choices and actions, and therefore, somewhere *else* in my head and in my heart, I feel in my bones that I am *not* a biochemical puppet or moist robot. When I am wearing my white lab coat

and doing experiments, or in my open-necked Brooks Brothers shirt and nice blue jeans teaching philosophy, I believe I am really a puppet and a robot; but when I get home, change into my comfortable T-shirt and baggy shorts, and open a beer, I believe I am really *not* a puppet or robot. This is philosophical *schizophrenia*, not a stable philosophical position.

More specifically, Pereboom claims that the "standard phenomenology of agency" is evidentially *neutral* as between Classical Libertarianism on the one hand, and Hard Determinism/Hard Incompatibilism/Soft Determinism on the other. That may be so *if* we construe the "standard phenomenology of agency" narrowly enough, but two things have gone seriously wrong here.

First, Pereboom has *not* considered the inherently richer, essentially embodied, agential phenomenology of vital sourcehood, natural creativity, and living in "natural open space."

Second, he has not considered, or even looked for, anything even remotely like non-classical Natural Libertarianism, but on the contrary has strategically boxed-in his metaphysical options by critically considering only the various versions of Classical Libertarianism as opposing doctrines.

So the plausibility and soundness of Pereboom's argument depends entirely on his having started with a needlessly poverty-stricken and boxed-in domain of phenomenological materials and metaphysical options. Therefore, to conclude this critical line of argument by modus tollens-ing Pereboom's modus ponens: *If* we do *not* start with these needlessly poverty-stricken and boxed-in materials, but instead on the contrary seriously consider the essentially embodied, agential phenomenology of vital sourcehood, natural creativity, and living in "natural open space," and also Natural Libertarianism, *then* Hard Incompatibilism is clearly false.

4.9 CONCLUSION

Here, then, is what this chapter shows us. *If*, as I have critically argued, Natural Mechanism, Hard Determinism, Soft Determinism, Classical Libertarianism (including classical and non-classical agent-causationism, non-causalism, and even-causal indeterminism), classical Compatibilism, classical Incompatibilism, Semi-Compatibilism, In-the-Zone Compatibilism, Revisionism, and Hard Incompatibilism are *all* false, *then* we must find a theory that explicitly rejects all of these doctrines *and also* conforms to our bedrock clear-and-distinct rational intuitions, grounded on self-evident phenomenology, about deep freedom, real human personhood, intentional agency, deep (non-)moral responsibility, and human rationality.

In my opinion, there is one and only one such theory: Natural Libertarianism. So in the next chapter, I want to complete my overall six-step argument for the truth of Natural Libertarianism by way of critically revisiting the recent and contemporary Harry Frankfurt-initiated debate about The Principle of Alternative Possibilities and moral responsibility.

Chapter 5

EITHER/OR: DEEP FREEDOM AND PRINCIPLED AUTHENTICITY

> The choice itself is crucial for the content of the personality: through the choice the personality submerges itself in that which is being chosen, and when it does not choose, it withers away in atrophy…. That which is to be chosen has the deepest relation to the one who is choosing, and when the choice is about an issue of elementary importance to life, the individual must at the same time continue to live, and this is why the longer he puts off the choice, the more easily he comes to alter it, although he keeps on pondering and pondering and thereby believes that he is really keeping separate the two alternatives of the choice.[351]

> My conception of freedom of the will appears to be neutral with regard to the problem of determinism. It seems conceivable that it should be causally determined that a person is free to want what he wants to want. If this is conceivable, then it might be causally determined that a person enjoys free will…. On the other hand, it seems conceivable that it should come about by chance that a person is free to have the will he wants. If this is conceivable, then it might be a matter of chance that certain people enjoy freedom of the will and that certain others do not. Perhaps it is also conceivable, as a number of philosophers believe, for states of affairs to come about in a way other than by chance or as the outcome of a [deterministic] sequence of natural causes. If it is indeed conceivable for the relevant states of affairs to come about in some third way, then it is also possible that a person should in that third way come to enjoy freedom of the will.[352]

5.0 INTRODUCTION

As advertised, in this chapter I will directly connect Natural Libertarianism to the recent and contemporary Harry Frankfurt-initiated debate about The Principle of Alternative Possibilities and moral responsibility, and in so doing, finish unpacking the internal structure of our essential free agency capacities for deep freedom, deep (non-)moral responsibility, and principled authenticity. Along the way, I will also argue that we have good reasons for retaining some carefully qualified features of each of the false standard views of classical Compatibilism, classical Incompatibilism, Hard Determinism, Soft Determinism, and Classical Libertarianism, by incorporating those carefully-qualified features into a distinctively different successor doctrine, namely, Natural Libertarianism. My conclusion will be that Natural Libertarianism

not only offers the best overall explanation of all the relevant empirical and a priori philosophical data about free agency, but also provides a cognitively, affectively, and practically liberating way of untying The Fourfold Knot of Free Agency. And that will complete the overall six-step argument for the truth of Natural Libertarianism that I previewed in section 1.3.

5.1 THE INTERNAL STRUCTURE OF DEEP FREEDOM

Many theories of free will—including Natural Libertarianism—hold that necessarily, I can freely choose or do *X only if* my choosing or doing *X*,

> (i) lacks an antecedent nomologically sufficient cause, and
> (ii) I myself am the causally sufficient ground, origin, or source of choosing or doing *X*.

The conjunction of these two necessary conditions is what Kant rather misleadingly calls "transcendental freedom," and what Robert Kane more aptly calls "ultimate responsibility."[353] But as I see it, these two necessary conditions provide only *part* of a single necessary but individually insufficient condition—namely, what I call *the real causal spontaneity condition*—for deep freedom, ultimate sourcehood, or up-to-me-ness. More precisely, according to Natural Libertarianism, the real causal spontaneity condition says that a necessary but individually insufficient condition of deep freedom and deep (non-)moral responsibility is that all my choosings and doings are

> causally efficacious in the physical world in a way that is fully consistent with, but also not entailed or otherwise necessitated by, the total set of deterministic or indeterministic causal natural laws, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially The Big Bang.

The satisfaction of this necessary condition (together with two further necessary conditions I will spell out below), in turn, ensure that a rational human agent or real human person comes to "enjoy" freedom of the will and deep (non-)moral responsibility in Frankfurt's "third way," that is, in a way that is equally distinct from those specified by Universal Natural Determinism and Universal Natural Indeterminism alike.

As I pointed out in chapter 3, according to Kant, transcendental freedom is how a human person can, "from itself" (*von selbst*) (*CPR* A533/B561), be the spontaneous mental cause of certain natural events or processes. If I am that human person, then insofar as I am transcendentally free, it follows that I am an ultimate source of my choices and intentional actions precisely because certain events or processes in physical nature are *up to me*—or to use Kant's own phrase, *in meiner Gewalt* (literally: "in my control" or "in my power"; *CPrR* 5: 94-95). So otherwise put, as I also put it in chapter 3, Kant's misleadingly-labeled transcendental freedom is in effect deep freedom of the will (aka ultimate sourcehood, up-to-me-ness), which in turn necessarily *includes* but also significantly *exceeds* the real causal spontaneity condition

alone. More specifically, it significantly exceeds the real causal spontaneity condition by adding

> (i) a condition requiring what I have called "the capacity for self-commitment to a live option," or the Kierkegaardian Either/Or, and also
> (ii) an *ownership* condition.

This completes a description of *the internal structure of deep freedom*, which I will discuss and defend in more detail shortly.

According to Natural Libertarianism, however, even the complex core metaphysical fact of deep freedom (= real causal spontaneity + the capacity for self-commitment to a live option or Kierkegaardian Either/Or + ownership) does not exhaust the *complete* fact of free will, which as I have said is more properly characterized as *free agency* in order to accommodate the further fact of *practical agency*. More precisely put, the free agency of rational animals or real persons has not only a complex structure, but more specifically an *hierarchical* or ordered-and-levelled, structure. Even more precisely still, according to Natural Libertarianism, the capacity for free agency has three logically distinct ordered sub-levels within it:

> (i) the capacity for veridical psychological freedom,
> (ii) the capacity for deep freedom (aka ultimate sourcehood, up-to-me-ness), and
> (iii) the capacity for principled authenticity.

The ordering is as follows. Veridical psychological freedom is a necessary but not sufficient condition of deep freedom. So (ii) requires (i), but also exceeds (i). Furthermore, deep freedom is a necessary but not sufficient condition of principled authenticity. So (iii) requires both (i) and (ii), but also exceeds both (i) and (ii). In this way, the three sub-levels of free agency are metaphysically *nested* like all the little Russian dolls inside the one big Russian doll; and correspondingly free agency is like the one big Russian doll—an inherently ordered, tri-leveled fact. Each of the nested sub-levels corresponds to a set of what I take to be basic authoritative rational intuitions[354] about free will, all grounded on self-evident phenomenology, so that each sub-level is a genuine *type* of free will—each sub-level is, as it were, its own little Russian doll. But only the complete, tripartite, ordered, leveled structure corresponds to all our basic authoritative rational intuitions about free will, because only the complete, tripartite, ordered, leveled structure fully captures how the several types of free will are internally necessarily related to one another in the total complex metaphysical fact of free agency. This total complex metaphysical fact of free agency, in turn, yields deep (non-)moral responsibility.

Therefore, according to Natural Libertarianism, *only* a theory of free agency that both logically distinguishes and also metaphysically nests or orders the three distinct sub-levels will be able to provide a philosophically adequate theory of free agency, and deep (non-)moral responsibility. All or at least most other classical and contemporary theories of free agency, no matter how clever and sophisticated, tend to be merely "flat," or one-leveled, as if only the psychological component on its own—standardly, also lacking the veridicality requirement—or only the deep freedom component on its own, or only the moral autonomy component on its own (+/- authenticity), could adequately account for free agency or deep (non-)moral responsibility.[355] Sometimes, of course, it is true that "less is more." But especially where

philosophical explanations are concerned, more often than not, on the contrary, it is true that *less is less* and *more is more.* Roll over and put away that razor, Ockham.

Back now to the condition of real causal spontaneity. What, more precisely, do I mean by this?[356] By *real causation,* I mean a metaphysically robust modal relation[357] between two singular spacetime events, e_1 and e_2, such that

> (i) e_2 is not earlier than e_1,
> (ii) e_1 nomologically sufficiently guarantees the existence and specific character of e_2, and
> (iii) e_2 would not have existed if e_1 had not existed.

Then e_1 is a nomologically sufficient real cause and e_2 is its effect. In turn, I will define *real causal efficacy* as follows:

> A singular spacetime event e_1 is really causally efficacious if and only if
>
> either (i) e_1 is itself a nomologically sufficient singular event cause of some spacetime event e_2, or
> (ii) e_1 is a necessary proper part of e_3, which itself is a nomologically sufficient singular event cause of e_2.

This notion of real causal efficacy can then be smoothly extended to properties and physical substances:

> A property *P* is really causally efficacious if and only if *P* is instantiated as an intrinsic relational, or immanent structural, property by events that are causally efficacious, and a physical substance *S* is causally efficacious if and only if *S* is constituted by causally efficacious events and properties.

And I define an event's *spontaneity* in the following way:

> An event *e* is spontaneous if and only if *e* is
>
> (i) causal-dynamically *unprecedented*, in the sense that *e* has never actually happened before,
> (ii) causal-nomologically *constrained-yet-also-underdetermined*, in the sense that *e* is fully consistent with, but also not entailed or otherwise necessitated by, the total set of deterministic or indeterministic causal natural laws, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or-energy facts about the past, especially The Big Bang,
> (iii) *creative*, in the sense of being recursively constructive, or able to generate infinitely complex outputs from finite resources, and
> (iv) *self-guiding*, in the sense of having an internally- or endogenously-driven teleology or purposiveness.

More generally, however, deep freedom (aka ultimate sourcehood, up-to-me-ness) includes three individually necessary, individually insufficient, and jointly sufficient conditions:

> (i) real causal spontaneity, as defined immediately above,
> (ii) the capacity for self-commitment to a live option, which is my ability to choose or do something *X*, as opposed to not choosing or doing *X*, and *X* would never actually happen if I were not to choose *X*, so that I can have an authentic choice in the Kierkegaardian sense of an Either/Or even if, in context, there are no alternative possibilities in the classical sense of branching futures and the future is temporarily not open, and
> (iii) ownership, which is the fact that that my choices and doings belong to me and my life, as a self-identical real person, and do not belong to some other agent or agency.

It is a striking feature of the contemporary metaphysical debate surrounding the problem of free agency that it is *exceptionally intuitionally polarized*, in a way that goes beyond all-too-familiar intuitional polarization in philosophy, in that those who engage in the free agency debate tend to find the prima facie "deep freedom intuition" *either* extremely compelling, with no room for compromise, *or else* a complete sham and wide open to a "debunking strategy" or "error-theory." This exceptional polarization of prima facie intuitions is revealing. Those who find it a complete sham and wide open to debunking and error-theorizing are, I think, reacting primarily to the further thought that the very idea of deep freedom automatically entails one or both of the two following theses:

> (i) rational agents or persons are special, unworldly substances causally operating outside of spacetime, and in such a way that nothing occurring in the series of events in nature prior to their choices or acts is in any way causally relevant to them (*classical agent causationism*), and/or
> (ii) a rational agent or person can always choose or act otherwise, in any possible set of circumstances, even if the entire causal history of the actual world prior to her choice or act were exactly replicated, and even if the alternative is completely irrelevant to the desires and beliefs of the agent, and this is a necessary condition of her causal and/or (deep) (non-)moral responsibility (*The Principle of Alternative Possibilities*, aka PAP).

But it is vividly clear to me that *both* classical agent causationism *and* PAP are false. So if either of them really *were* entailed by the very idea of deep freedom, then I too would find the deep freedom intuition a complete sham. But I don't find the deep freedom intuition a complete sham—contrariwise! So classical agent causation and PAP really *ain't* entailed by it. That's logic.

In other words, according to Natural Libertarianism, deep freedom entails neither classical agent causationism nor PAP. Indeed, when I say that a choice or act is deeply free, I am also assuming that

> (i) many things that happened in physical nature prior to my choices or acts are not only causally relevant to my choices or acts, but also causal-nomological constrainers[358] of them,
> (ii) necessarily, all my choices and acts occur in actual spacetime and in the actual event-sequence, and

> (iii) I can freely and deeply (morally or non-morally) responsibly choose or do *X*, or refrain from so choosing or so doing, without either my believing in, or there in fact being, in that context, any distinct possible alternative *Y* in the classical sense of indeterministic branching futures, so that the future is temporarily not open.

If (i) were not true, then all my choosings and doings would be alienated from my own earlier real personal life history in the physical natural world, and would thus fail to belong to me and me alone. So (i)'s falsity would violate not only the second condition on the spontaneity of deep freedom, but also the ownership condition.

In turn, if (ii) were not true, then none of my free choices and acts could be causally efficacious, unless they were non-standardly causally overdetermined. So (ii)'s falsity would also violate the real causal spontaneity condition on deep freedom.

And if (iii) were not true, then I could never freely and deeply (non-)morally responsibly choose or do things single-mindedly or when the chips are already down. So (iii)'s falsity would violate the capacity for self-commitment to a live option condition of deep freedom.

But it is perfectly consistent with (i) and (ii) that

> (iv) none of the causally relevant and causal-nomologically constraining prior natural physical things mentioned in (i) is either individually or collectively nomologically causally sufficient for either the existence or specific character of my choice or act, so the total set of such things not only constrains but also underdetermines my choice or act, and
> (v) I myself, and not anything or anyone else, am the real causal spontaneous nomologically sufficient cause of my choice or act.

Therefore (i) and (ii) are both perfectly consistent with my being the efficacious natural cause of all my choosings or doings and my ownership of all those choices or acts.

Moreover, it is perfectly consistent with (iii) that

> (vi) I can choose or do *X*,
> (vii) I might have not chosen or not done *X* by just failing or refusing to choose *X* (I will call these *the null choice options*),
> (viii) *X* would never actually happen (or: would not have happened) if I were not to choose (or: if I had not chosen) *X*, and
> (ix) conditions (vi) through (viii) can obtain even if, in context, I have no alternative possibilities in the classical sense of branching futures, and the future is temporarily not open.

Or in other words, if (vi), (vii), (viii), and (ix) are all true, then even if there are in fact no possible alternatives *Y*, in that context, to my choosing or doing *X*, or not choosing or doing *X*, it still is the case that both the existence and specific character of *X* are necessarily dependent on me, and *X* is still *a live option* for me. In other words, I can make *X* happen by choosing it, and *X* would never actually happen if I were not to choose it. This fact, yet again, is what I call "the capacity for self-commitment to a live option" or "the Kierkegaardian Either/Or."

Granting me for the purposes of argument at least the *intelligibility* of the idea of the capacity for self-commitment to a live option or the Kierkegaardian Either/Or, now what I want

to argue is that *the dynamacist model of life* that I developed in chapter 2, together with *the biologically-based theory of practical agency* that I developed in chapter 3, together with *the Incompatibilistic Compatibilism* that I developed in chapter 4, jointly provide an adequate explanation and metaphysical foundation for the real fact of deep freedom that also centrally includes the capacity for self-commitment to a live option or Kierkegaardian Either/Or, as embedded within the larger metaphysical framework of the agency-constituting capacity for principled authenticity.

Fully explicitly now, according to Natural Libertarianism, deep freedom is the conjunction of three individually necessary, individually insufficient, and jointly sufficient conditions, as follows.

> A rational animal or real person *P*'s choosing or doing *X* is deeply free if and only if
>
> (i) *P* is the far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned, thermodynamic, causal-dynamically unprecedented, causal-nomologically constrained-yet-also-necessarily-underdetermined, creative, and self-guiding efficacious cause of choosing or doing *X*, or refraining from so choosing or so doing (*that is, the real causal spontaneity condition is satisfied*),
> (ii) *P* can choose or do something *X*, as opposed to not choosing or doing *X*, and *X* would never actually happen (or: would not have happened) if *P* were not to choose (or: had not chosen) *X*, so that *P* has a live option for self-commitment in the Kierkegaardian sense of an Either/Or, even if, in context, there are no alternative possibilities in the classical sense of indeterministic branching futures, and the future is temporarily not open (*that is, the capacity for self-commitment to a live option condition is satisfied*), and
> (iii) *P*'s choosings and doings belong to her and her life, as a self-identical real person, and do not belong to some other agent or agency (*that is, the ownership condition is satisfied*).

If all this is correct, then the fundamental problem with other classical or contemporary theories of free will is that they have failed to recognize these two crucial points:

> (i) free agency is freedom-in-life, and
> (ii) the nature of deep freedom, the nature of rational (human) animality, and the nature of real (human) personhood, are all ultimately one and the same.

An extremely important and theoretically fruitful feature of Natural Libertarianism is the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, that is built into the larger fact of deep freedom. As I have noted above many times, if I have the capacity for self-commitment to a live option, then

> (i) I can choose or do *X*,
> (ii) I can not-choose or not-do *X* by just failing or refusing to choose *X* (= the null choice options),
> (iii) *X* would never actually happen (or: would not have happened) if I were not to choose (or: if I had not chosen) *X*, and

> (iv) conditions (i) through (iii) can obtain even if, in context, I have no alternative possibilities in the classical sense of branching futures, and the future is temporarily not open.

The first and second clauses jointly express a metaphysically robustly or really possible actual-sequence option for the agent, namely, "the live option." The third clause expresses a metaphysically robust robust counterfactual causal condition. And the fourth clause says that my having the ability for self-commitment with respect to *X* entails that I can go-for-*X*, or not go-for-*X*, and necessarily *X* would not happen in the actual world without me and my choice to go-for-*X*, even if, in the actual world in that very context, I have no alternative possibilities and my context-sensitive future is not open. Or in other words, alternative possibilities in the classical sense of indeterministic branching, open futures are inherently *irrelevant* to deep freedom.

It is crucial to note, however, that this fourth clause does *not* claim that there are *no such things as classical indeterministic alternative possibilities*, in the sense of branching, open futures. On the contrary, *there certainly could be classcal indeterministic alternative possibilities*, in the sense of branching, open futures, as a matter of logical-conceptual, "weak metaphysical," or analytic possibility, as a matter of non-logical, "strong metaphysical," real, or synthetic possibility, or as a matter of nomological or physical possibility. And not only that, *there really are classical indeterministic alternative possibilities*, in the sense of branching, open futures, *wherever* there is some significant microphysical or macrophysical indeterminism in nature. None of these is ruled out by Natural Libertarianism. Indeed, quantum physics strongly indicates the existence of various kinds of indeterministic facts in nature—and I am accepting this as an empirical truth. It is just that such indeterministic facts are inherently *irrelevant* to free agency, *except* insofar as they factually break the classical metaphysical slave-chains of Universal Natural Determinism and thereby minimally entail the truth of *non*-determinism.

So what metaphysically matters for free agency is *not* metaphysical indeterminism. In order for there to be a fully intelligible, defensible, and true version of Libertarianism, namely, a fully intelligible, defensible, and true *Natural* Libertarianism, all that is essentially required are:

> (i) the truth of non-determinism—that is, the truth of the denial of Universal Natural Determinism—as, for example, minimally entailed by various factual truths of quantum physics,
> (ii) the truth of anti-mechanism, that is, the manifest reality of natural purposiveness or teleology, and
> (iii) the real causal spontaneity of minded animals, including of course rational human animals.

Of course, the belief in the normative and metaphysical significance of branching, open futures and alternative possibilities matters quite a lot from the third-person standpoint of (neo)liberal democratic political theory and political libertarianism. According to liberalism (whether classical liberal or neoliberal) and also according to political libertarianism, in a world in which people are inherently self-interested and mutually antagonistic, and have collectively entered into some sort of social contract that legitimates a (maximal, moderate, or minimal) government authority, incorporating a territorial monopoly on the power to coerce the people

living in that territory and governed by that authority, in order to keep the peace and make the mutual pursuit of self-interest possible for everyone, then it is significantly better for people to have lots of different things available for them to choose, than for them to have fewer options. And correspondingly, according to (neo)liberals and political libertarians, excessively authoritarian, regressive political regimes that antecedently shut down or arbitrarily narrow the total class of socially-available alternatives are radically worse than those less authoritarian regimes that permanently sustain or progressively open up the total class of socially-available alternatives. These (neo)liberal and political libertarian truisms—usually accompanied by the mental imagery of a flag gently waving in the wind, the stirring strains of a national anthem, and rows upon rows of people with their right hands held over their hearts and glassy, staring eyes—no doubt provide a good sociocultural explanation for why some or another version of Classical Libertarianism has seemed to many 20th and 21st century Anglo-American philosophers of free will to be *the only game in town*.[359]

As I mentioned in chapter 3, the notion of exercising the capacity for self-commitment to a live option, or Kierkegaardian Either/Or—whereby, at some time or another, I manifest or realize the capacity to exercise or not exercise any *other* ability that I might have—is closely related to the fundamental notion of "trying" in action-theory, and more precisely is identical to what Maiese and I have called "effortless trying."[360] Effortless trying is pre-reflective and conscious, but *non-self-consciously* conscious, effective desiring, and therefore need not involve any kind of self-conscious or reflective effective desiring. In turn, effortless trying is presupposed by all self-conscious or reflective, effortful trying. So given my self-commitment to *X*, if *X* actually happens, then *X* happens because I either effortlessly (pre-reflectively) tried or effortfully (self-consciously or reflectively) tried to do *X*, and because my trying made *X* actually happen.

There are also some obvious similarities between the notion of self-commitment to a live option and the conception of freedom found in various versions of Existentialism. In particular, as I have stressed, the notion of a capacity for self-commitment to a live option is closely related to Kierkegaard's profound doctrine of an Either/Or-driven self-choosing that pre-reflectively consciously characterizes every moment of a particular conscious, intentional, caring, rational animal's or real person's life, no matter what universal rational or natural laws or other external constraints there might be. The notion of the self-commitment option is also closely related to Sartre's notion of a "radical choice" that is perfectly consistent with cases in which the chips are already down—when *les jeux sont faits*.[361]

Sartre is sometimes interpreted as having proposed an extreme version of classical or non-classical agent causationist Classical Libertarianism—but I think that this is a serious misinterpretation of his view. In any case, and quite apart from vexed historical-interpretive questions, the basic shared Kierkegaardian and Sartrean existentialist idea here, as I understand it, is that what is essential to something's being deeply free and deeply (non-)morally responsible, is that I be able to choose or do something, as opposed to not choosing or not doing that thing, such that the end or goal of my choice or act would never actually happen (or: would not have happened) were I not to choose it (or: had I not chosen it). And that it is therefore ultimately metaphysically irrelevant whether, in that context, I have any classical indeterministic branching alternatives, and thus metaphysically irrelevant whether the future is open for me right then and there.

One foreseeable objection to my view is that, even despite what I have already argued, the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, "must be" a *special*

kind of classical alternative possibility, which would then implicitly reinstate The Principle of Alternative Possibilities, aka PAP. But this objection involves a strategically uncharitable and false interpretive presupposition—namely, that *all agency-relevant optionality must be non-actual-sequence possibility, namely, indeterminism*. Dialectically, this is highly analogous to claiming that every non-physicalist position in the philosophy of biology "must be" a version of substance vitalism; and also highly analogous to claiming that every non-physicalist position in the philosophy of mind "must be" a form of substance dualism. The underlying *bad philosophical picture* that is built into both of these pairs of "must-be" or "obvious" options, is this: that the logical space of possible solutions to the problem of life and to the mind-body problem is exhausted by the simplistic binary architecture—

PHYSICALISM-OR-DUALISM.

In the philosophy of free will, essentially the same underlying bad philosophical picture pre-establishes, without argument, that the logical space of solutions to the free will problem is exhausted by the equally simplistic binary architecture—

DETERMINISM-OR-INDETERMINISM.

Nevertheless, just as at least one fully intelligible and defensible *non*-physicalist, *non*-dualist positions are available in the philosophy of biology (for example, dynamicism) and in the philosophy of mind (for example, the essential embodiment theory), so too at least one tenable *non*-determinist, *non*-indeterminist position is available in the metaphysics of free will: Natural Libertarianism. So the non-determinism built into Natural Libertarianism does *not* entail PAP.

The classical notion of an alternative possibility is that, relative to some context, the past and the causal laws of nature can remain exactly the same, and then one could choose or do either *X* or *Y*, even if *Y* is completely irrelevant to the desires and beliefs of the agent. More precisely, as we saw above, the classical notion of alternative possibilities says this:

> A person can always choose or act otherwise, even if the entire causal history of the actual world prior to her choice or act were exactly replicated, and even if the alternative is completely irrelevant to the desires and beliefs of the person, and this is a necessary condition of her causal and/or (deep) (non-)moral responsibility.

So according to the classical notion of alternative possibilities, the future is open in that it has definite branches (aka "open doors"), and in that context you could have chosen or acted either way, no matter what the past and laws of nature were, and no matter whether the alternative actually matters to you. Such alternatives might be metaphysically available as a matter of logical, analytic, or conceptual possibility (aka weak a priori metaphysical possibility), assuming that Fatalism is false; or they might be metaphysically available as a matter of nomological or physical possibility (aka weak a posteriori metaphysical possibility, that is, logical possibility plus some factual "meaning postulates" about causal natural laws), assuming, in line with quantum mechanics, that some form of indeterminism is factually true about at least some parts of physical nature. But in either case they would be at most

metaphysically *non-robust* possibilities. That is, they would *not*, in and of themselves, be what we can choose or do in *the actual event-sequence*.

But what is at issue in the capacity for self-commitment to a live option or Kierkegaardian Either/Or is *only* whether, in that context, the choice to pursue *X* will be made by you, so that *X* can come into existence in the actual sequence because of your trying, or whether *X* will fail to exist in the actual sequence for lack of your trying. According to Natural Libertarianism, then, *only the actual sequence and the actual world really and truly matter* to deep freedom of the will/deep (non-)moral responsibility and free agency. Thus it is only the set of metaphysically robust, non-logical, essentially non-conceptual, real or synthetic (aka strong metaphysical) possibilities in the actual sequence and in the actual world, that is, the possibilities that are supplied by the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, that really and truly matter to deep freedom of the will/deep (non-)moral responsibility and free agency.

Now this ability for self-committing-to-a-live-option, or Kierkegaardian Either/Or, can obtain *even if only X was ever possible in the actual world*, due to the actual contextual presence of a special set of what I will call "causal-interventionist counterfactual conditions," aka CICCs (pronounced like 'kicks'). Given the existence of some CICCs, it is true that even if you *were* to try to choose or do something other than *X*—which of course you actually do not—then those CICCs *would* ensure that *X* happens, even though you did not freely choose or do *X*.

It is very important to see that the actual contextual presence of some CICCs is sharply different from the actual contextual presence of any special set of conditions under which the intentional agent's particular choice or act is inherently naturally determined, whether distally or proximally, or else caused by a powerful intervener, say, a Cartesian evil demon or mad/malicious scientist. Otherwise put, the concept of CICCs is *not* the same as the concepts of either Universal Natural Determinism or "manipulation." This is because the actual contextual presence of some CICCs is perfectly consistent with self-commitment to a live option and therefore also with deep freedom in the actual event-sequence, whereas Universal Natural Determinism and powerful manipulation are both fully *inconsistent* with self-commitment to a live option and deep freedom/deep (non-)moral responsibility in the actual sequence. The latter point about full inconsistency is shown by my arguments for local incompatibilism with respect to Natural Mechanism in section 4.5 above. In any case, once we have made this conceptual distinction between CICCs on the one hand, and Universal Natural Determinism/powerful manipulation on the other, then we should be able to see clearly that the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, is essentially unaffected by all of the (perhaps, all-too-familiar) contemporary "Frankfurt-style" counterarguments against PAP.

In the next section, I will elaborate, and then argue explicitly for, that last claim—namely, that the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, is *essentially unaffected* by all of the contemporary Frankfurt-style counterarguments against PAP. If successful, this argument will demonstrate that Frankfurt-style counterarguments against PAP are smoothly consistent with local incompatibilism with respect to Natural Mechanism, and to that extent, they are also sharply inconsistent with classical Compatibilism. The philosophical moral of this story is that you do not have to be a classical compatibilist in order to accept Frankfurt-style counterarguments against PAP. You can consistently accept Frankfurt-style counterarguments against PAP *and* still be a thoroughly non-classical kind of incompatibilist. That is, you can still be an "incompatibilistic compatibilist" and a Natural

Libertarian. It is, therefore, a big mistake to think that the acceptance of Frankfurt-style counterarguments against PAP somehow rationally forces classical Compatibilism/Soft Determinism on us, even if, as a matter of sheer philosophico-sociological fact, almost all of the contemporary defenders of Frankfurt-style counterexamples to PAP are *also* defenders of classical Compatibilism/Soft Determinism. Obviously, however, *the philosophical majority can be completely wrong*—even if the personal, social, and ideological pressures of professional academic life sometimes make this obvious point very hard to remember.

5.2 FROM FRANKFURT BACK TO KIERKEGAARD: HOW TO HAVE A LIVE OPTION, OR KIERKEGAARDIAN EITHER/OR, *WITHOUT* ALTERNATIVE POSSIBILITIES

In his seminal 1969 paper, "Alternate Possibilities and Moral Responsibility," Harry Frankfurt offered a famous argument against PAP, which he defined as follows:

> This principle states that a person is morally responsible for what he has done only if he could have done otherwise.[362]

Frankfurt's argument mainly consists in providing an intuitive counterexample to PAP, involving a conceivable set of causal-interventionist counterfactual conditions, namely, CICCs, which can then be easily generalized to a class of "Frankfurt-style counterexamples":

> Suppose someone—Black, let us say—wants Jones_4 to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones_4 is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones_4 is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones_4 is going to decide to do something else, Black takes effective steps to ensure that Jones_4 decides to do, and that he does do, what he wants him to do. Whatever Jones_4's initial preferences and inclinations, then, Black will have his way…. Now suppose that Black never has to show his hand because Jones_4, for reasons of his own, decides to perform and does perform the very action Black wants him to perform. In that case, it seems clear, that Jones_4 will bear precisely the same moral responsibility for what he does as he would have borne if Black had not been ready to take steps to ensure that he do it…. <u>This, then, is why the principle of alternate possibilities is mistaken. It asserts that a person bears no moral responsibility—that is, he is to be excused for having performed an action, if there were circumstances that made it impossible for him to avoid performing it. But there may be circumstances that make it impossible to avoid performing an action without those circumstances in any way bringing it about that he performs the action.</u>[363]

So PAP is false.

Sometimes it is argued that even if PAP is false, nevertheless there is still a weakened version of it that could be accepted by defenders of classical Compatibilism/Soft Determinism,

which depends on the idea that this following counterfactual could be true in a universally naturally deterministic world:

> Had a person wanted to choose or do otherwise, then she could have chosen or done otherwise.

In other words, this spells out a specifically counterfactual-conditional notion of alternative possibilities, which leads on naturally to a counterfactual-conditional version of PAP:

> The Principle of Alternative Possibilities^{cc}, aka PAP^{cc}: A person is morally responsible for what she chooses or does only if, had she wanted to choose or do otherwise, then she could have chosen or done otherwise.

What about PAP^{cc}? The problem is that *it* has some Frankfurt-style counterexamples *too*. Suppose that someone is a "willing drug addict" who is fully self-committed to her addiction, and also fully addicted.[364] It is true of her now that even if she had wanted to stop taking the drug, she could not stop taking the drug. Her drug-taking choices and acts are causally overdetermined by her full addiction. So she does not even have counterfactual-conditional alternative possibilities with respect to taking the drug. Yet she is (deeply) morally responsible for her addiction, for better or for worse. So PAP^{cc} is false.

Frankfurt-style counterexamples to PAP and PAP^{cc}, it might seem, are fully consistent with the truth of Universal Natural Determinism, and also, it might seem, are fully consistent with the truth of either a partial version of Natural Indeterminism or Universal Natural Indeterminism. So someone whose prima facie intuitions tell him about these apparent consistencies might well conclude from Frankfurt-style counterexamples that (deep) (non-)moral responsibility is fully consistent with Universal Natural Determinism and with Natural Mechanism, and that either Compatibilism/Soft Determinism, Semi-Compatibilism, or Revisionism is true.

But both of these conclusions would be non sequiturs. Frankfurt-style counterexamples to PAP always postulate the actual contextual presence of some special set of causal-interventionist counterfactual conditions, namely, of some CICCs, or another. But the actual contextual presence of some CICCs does not in fact entail Universal Natural Determinism or powerful manipulation in the actual event-sequence. So the presence of some CICCs does not in fact entail Compatibilism/Soft Determinism.

In this connection, David Widerker[365] and others have correctly pointed out that Frankfurt-style counterexamples cannot work without presupposing the existence of some uncompelled and unmanipulated volitional powers in the agent (in the original example, "Jones$_4$"). That is because the intervening manipulator (in the original example, "Black") cannot intervene until the agent has started to deviate (or, as per some post-Frankfurtian examples, until some mechanism detects the neurobiological beginnings of such a deviation) from the manipulator's plan, in the counterfactual sequence of events. In the actual sequence of events, by contrast, the intervening manipulator never has to intervene, because, as Frankfurt puts it (with underlining added):

> Jones$_4$, for reasons of his own, decides to perform and does perform the very action Black wants him to perform.

It is assumed by Widerker and many others that these uncompelled and unmanipulated powers of the agent must also be *indeterministic*, and that they therefore, somehow or another, reinstate PAP. Of course, we already know that indeterminism deeply threatens free will too; and like Frankfurt I am also deeply skeptical about PAP and PAP^{cc} alike. So I think that it is a serious mistake to claim that the uncompelled and unmanipulated powers of the agent are indeterministic. Correspondingly, I want radically to strengthen Widerker's objection in such a way that it entails *neither* the oxymoronic notion of indeterministic free will, *nor* PAP, *nor* PAP^{cc}. According to Natural Libertarianism, then, the uncompelled and unmanipulated powers of the agent are *not* what Fischer and others have called an indeterministic "flicker of freedom." Instead and sharply on the contrary, these powers are a far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, finegrainedly normatively attuned thermodynamic *growth spurt* of non-deterministic, non-indeterministic, naturally purposive or teleological, naturally creative deep freedom, in the specific form of an actual exercise of the capacity for self-commitment to a live option, in the actual sequence of events.

More precisely, it is absolutely clear that all Frankfurt-style counterexamples to PAP *presuppose* that agent has already self-committed to a live option, whether in the form of a decision *X* or course of action *X*, *in the actual sequence*, when, again,

> for reasons of his own, [the agent] decides to perform and does perform the very action [the counterfactual intervening manipulator] wants him to perform.

This is because the counterfactual intervening manipulator cannot intervene until it is also a fixed element *in the counterfactual sequence alone* that the agent has self-committed to a live option *Y* rather than to *X*, and that it is *X* and not *Y* that the counterfactual intervening manipulator wants. The counterfactual intervening manipulator then brings it about *in the counterfactual sequence alone* that *X* is the only thing that the person can choose or do, which shows that PAP is false. So Frankfurt-style counterexamples all presuppose our capacity to self-commit to a live option *in the actual sequence*, and thus they all presuppose the Kierkegaardian Either/Or, even if, because of the actual contextual presence of a special set of causal-interventionist counterfactual conditions, some CICCs, we could not have chosen or acted in a way that is different from what we are already self-committed to.

Therefore, in Frankfurt-style counterexamples to PAP, at the very least, the rational animal or real person, the free agent, always has the live option of choosing *X* or not choosing *X* (according to either of the null choice options of failing to choose or refusing to choose) such that *X* would never happen (or: would not have happened) in the actual sequence if the agent were not to choose (or: had not chosen) *X*, even if, in context, she lacks any alternatives in the classical sense of branching futures, and the future is temporarily not open. Again, the uncompelled and unmanipulated powers of the agent in Frankfurt-style cases are nothing more and nothing less than the non-deterministic, non-indeterministic growth spurt, in the actual sequence, of the agent's exercise of the constantly present capacity for self-commitment to a live option, or Kierkegaardian Either/Or, that inherently belongs to deep freedom.

Similar points can be made about Martin Luther's famous expression of his freely willed and almost paradigmatically (deeply) morally responsible choice of religious iconoclasm, often cited as further support for Compatibilism, for example, by Dennett:[366] "Here I stand. I can do no other." It is true that, on the face of it, this sounds like Luther is asserting the compatibility of moral responsibility and Universal Divine Determinism. But on deeper and further

reflection, and refusing to beg the question—notice that Luther did *not* say: "Here I stand, I can do no other *than God made me do*"—we can recognize that the best interpretation of what Luther is saying, *is that it expresses a proto-Existentialist self-commitment to his live option of religious iconoclasm*. On this interpretation, Luther is saying that if he had either failed or refused to choose this live option (= the null choice options), *then he would have been untrue to himself, that is, inauthentic*. Or as Kierkegaard would put it, Luther would have failed to "choose himself." So on this interpretation, it is clearly not correct to say that Luther had no options at all, or that he was causally determined. On the contrary, he might have *not* chosen and done what he actually chose and did, according to either of the null choice options, and thereby he might have failed to live up to his own highest moral principles and values. So, on this interpretation, Luther fully exercised his capacity for self-commitment to a live option, that is, he fully engaged in a Kierkegaardian Either/Or, *even if*, in that context, he lacked any alternative possibilities in the classical sense of indeterministic branching futures and the future was temporarily not open.

—Due, for example, to the actual contextual presence of some CICCs created by an omniscient, omnipotent, and omnibenevolent Protestant God, while S/he was briefly taking a rest on Sunday from Her/His actual-sequence Divine Determinist labors, by instead, in this context, merely playing the role of Frankfurt's counterfactual intervener, Black, in order to guarantee Luther's compliance to the dictates of *his own* conscience. Therefore, all non-question-begging Luther-type counterexamples to PAP *also* presuppose the capacity for self-commitment for a live option, or Kierkegaardian Either/Or.

But then it follows that it is the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, as a constituent condition of deep freedom, together with the other constituent conditions of real causal spontaneity and ownership, as metaphysically embedded in a larger structure that includes the capacity for principled authenticity, that is the collectively necessary and sufficient condition of (deep) (non-)moral responsibility in the Frankfurt-style and non-question-begging Luther-type examples, and in all other cases of (deep) (non-)moral responsibility too. As Fischer has aptly and indeed profoundly put it, the "moral" of the Frankfurt-style counterexamples is that *if* Universal Natural Determinism is incompatible with moral responsibility, then this incompatibility is *not* because the intentional agent lacks alternative possibilities.[367] The defender of Natural Libertarianism will fully agree with the truth of this negative conditional statement. But it does not follow from this true negative conditional claim, as Fischer thinks, that either Compatibilism or Semi-Compatibilism is true. So what the defender of Natural Libertarianism will then say in response to Fischer's profoundly apt point is this:

> "You say that the moral of the Frankfurt-style counterexamples is that *if* Universal Natural Determinism is incompatible with moral responsibility, then this incompatibility is *not* because the intentional agent lacks alternative possibilities. That is absolutely right, although not in the way you intended. For, contrary to what you yourself think, the antecedent of your conditional is true. Universal Natural Determinism is locally incompatible with (deep) (non-)moral responsibility. This, in turn, is precisely because Universal Natural Determinism is locally incompatible with deep freedom with respect to Natural Mechanism. And this, in turn, is ultimately because Universal Natural Determinism is locally

> incompatible with the online capacity for self-commitment to a live option, or Kierkegaardian Either/Or."

Otherwise put, the notion of the capacity for self-commitment to a live option, or the Kierkegaardian Either/Or, which according to Natural Libertarianism captures a necessary proper part of deep freedom, is not metaphysically consistent with Universal Natural Determinism, precisely because local incompatibilism with respect to Natural Mechanism is true, as I argued in section 4.5 above, and also because deep freedom requires real causal spontaneity.

Fischer's "deep control," for all its epistemic depth, is nevertheless, from the point of view of Natural Libertarianism, metaphysically shallow and normatively hollow, yielding at best *shallow, hollow moral responsibility*, and neither deep moral responsibility nor deep non-moral responsibility. But sharply on the contrary, we are, for better or worse, *the creative natural artists of our own lives*, and deeply (non-)morally responsible for them, not *their crafty poker-players*, like Sinatra and his pals in the Rat-Pack movies, "doing it my way." So Frankfurt-style and Luther-type counterexamples do *not* demonstrate the compatibility of (deep) (non-)moral responsibility with Universal Natural Determinism, even if they *do* indeed effectively undermine PAP and PAP^{cc}. That is the second and *deepest* moral of the Frankfurt-style counterexamples.

Again, and now to conclude this part of the discussion, according to Natural Libertarianism, it is deep freedom, as essentially including the capacity for self-commitment to a live option, or the Kierkegaardian Either/Or, as embedded in a larger metaphysical structure that includes the agency-constituting capacity for principled authenticity, that is necessary and sufficient for deep (non-)moral responsibility. Alternative possibilities are both metaphysically irrelevant and epistemically irrelevant for free agency—even if they do remain epistemically relevant for classical liberal, neoliberal, or libertarian politics. The crucial mistake made by defenders of PAP and PAP^{cc} alike was to confuse the genuine necessity, for deep freedom and deep (non-)moral responsibility, of *actual sequence optionality*, with the illusory necessity of an array of indeterministic branching, open futures, aka "open doors." As Locke's famous example shows, and Frankfurt-style counterexamples reinforce with modal-metaphysical subtlety, there might be

> either (i) no open doors in the actual sequence (Locke),
> or else (ii) doors that wouldn't open *only if* you were to change your mind and want to leave the room (Frankfurt).

So *alternative possibilities are altogether otiose for free agency*: all that is required and sufficient for deep (non-)moral responsibility is at least one live option and the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, when it is embedded in the larger free-agency-constituting metaphysical structure of deep freedom and principled authenticity.

5.3 PSYCHOLOGICAL FREEDOM, DEEP FREEDOM, AND PRINCIPLED AUTHENTICITY

As I mentioned at the beginning of this chapter, according to Natural Libertarianism, the free agency of rational animals or real persons has three logically distinct but also metaphysically nested or ordered levels embedded in it:

> (i) the capacity for veridical psychological freedom,
> (ii) the capacity for deep freedom, and
> (iii) the capacity for principled authenticity.

And according to Natural Libertarianism, as I also mentioned, strictly speaking, psychological freedom, when it is unspecified as to whether it is veridical or non-veridical and formulated as a weak disjunction—"veridical-or-non-veridical psychological freedom"—is also a necessary but not sufficient condition of deep freedom; and deep freedom, ultimate sourcehood, or up-to-me-ness, is a necessary but not sufficient condition of principled authenticity. Let us now look more closely at the three levels of the free agency structure.

The first level of the free agency structure is veridical psychological freedom. Now psychological freedom, per se, without regard to its veridicality or non-veridicality, is my first-order consciousness[368] of being both negatively and positively free. Or otherwise put, psychological freedom, per se, is my subjective experience of having an unfettered and really causally spontaneous will. This consciousness, in turn, can be

> either (i) a correct or true consciousness of being both negatively and positively free, such that it is an actual fact that I am both negatively and positively free, in which case, it is what I call *veridical psychological freedom*, so that at any time, the intentional subject is either in one kind of state or else in the other, never both, or (ii) an incorrect or false consciousness of being both negatively and positively free, that is, a mere seeming to be both negatively and positively free, such that I in fact am neither negatively nor positively free, in which case, it is what I call *non-veridical psychological freedom*.

Correspondingly, I am also committed to the thesis of *strong metaphysical disjunctivism* about the difference between veridical psychological freedom and non-veridical psychological freedom,[369] which says that

> (iii) veridical psychological freedom and non-veridical psychological freedom essentially share no intentional or phenomenological content whatsoever, even if they accidentally share some (or even many) other psychological or non-psychological properties, and
> (iv) the difference between veridical psychological freedom and non-veridical psychological freedom is, in principle, inherently *discriminable* for rational human agents, even if, in context, or in a specific range of contexts, it is actually *undiscriminated* because the agent's capacity for discrimination is adversely affected or suppressed in that context or those contexts.

So obviously psychological freedom, per se, when it is *non*-veridical psychological freedom, is consistent with my *not* really being negatively or positively free. Furthermore, non-veridical psychological freedom is also inherently conceptually-determined, that is, inherently open to "cognitive penetration." By sharp contrast, *veridical* psychological freedom is essentially non-conceptual,[370] hence it is also pre-reflectively conscious, or non-self-conscious and non-reflective, and therefore impervious to "cognitive penetration." But since rational human agents like us are *also* self-conscious and reflective, then the possession of psychological freedom for us, whether it is veridical psychological freedom or non-veridical psychological freedom, also entails a further capacity for having self-directed beliefs to the effect that we are negatively and/or positively free.[371] In the case of veridical psychological freedom, obviously, those self-directed beliefs that we are negatively and/or positively free, are *true* beliefs; whereas in the case of non-veridical psychological freedom, they are *false* beliefs. Thus we can be deceived about our freedom and have what I call *sheer illusions* of deep freedom.

These sheer illusions can occur in dreams, hallucinations, or most poignantly, in causally-overdetermined pathological waking psychological states, such as cases of sociopathic paranoid schizophrenics who believe they are choosing and acting with free will, but are actually delusional and insane, and acting under an irresistible compulsion when they commit crimes, hence are correctly legally judged to be "not guilty by reason of insanity." One such case, or so it seems from the available evidence, is the real-world waking nightmare of a former Yale Law School student, Ketema Ross:

> Early one morning in 2007, Ross heard President George W. Bush [Yale] '68 telling him that his next-door neighbors were traitors who needed to be gotten rid of. Ross broke into the elderly couple's apartment and beat them with a broom handle. (They both survived the attack.) Charged with assault, he pleaded not guilty by reason of insanity. Now Ross says he has recovered his sanity, and a court order says he is no longer "a substantial danger." And, after seven years of confinement in a psychiatric hospital, he has regained his freedom, mostly: by court order, he was conditionally discharged on January 11 [2015].[372]

The second level of the free agency structure is deep freedom, which, as we saw above, has the following analysis: A rational animal or real person *P*'s choosing or doing *X* is deeply free, flows from her ultimate sourcehood, or is up to her if and only if

> (i) the real causal spontaneity condition is satisfied by *P*,
> (ii) the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, condition is satisfied by *P*, and
> (iii) the ownership condition is satisfied by *P*.

Now psychological freedom, *per se*, is not a sufficient condition of deep freedom, precisely because, as some philosophers (including Spinoza, Leibniz, Hume, and more recently Frankfurt,[373] and many other Frankfurt-inspired philosophers of agency) have correctly pointed out, both the pre-reflective consciousness, and also the self-conscious or reflective belief, of having an unfettered and spontaneously really causal will, and of being both negatively and positively free, are perfectly consistent with Universal Natural Determinism. Formulated in my

terminology, if Universal Natural Determinism (or, indeed, Natural Mechanism) were true, then all of these conscious or self-conscious states would be cases of *non*-veridical psychological freedom and/or false self-directed beliefs about my deeply free will.

Nevertheless, both veridical psychological freedom and also veridical psychological freedom's discriminability from non-veridical psychological freedom, collectively yield a conjunctive necessary epistemic condition of deep freedom, by way of the capacity for self-commitment to a live option, or Kierkegaardian Either/Or. This conditions says that no one could have a capacity for self-commitment to a live option X and at the same time

> either (i) be in a state of non-veridical psychological freedom,
> or (ii) truly believe herself to be prevented from choosing X or doing X,
> or (iii) truly believe herself to be inwardly or outwardly compelled to choose X or do X,
> or (iv) truly believe herself to be otherwise unable to choose or do what she wants.

Correspondingly, by *a self-consciously or reflectively fettered, epiphenomenal, will*, I mean one's self-conscious or reflective awareness to the effect that, and correspondingly one's belief about oneself to the effect that, one is helplessly violated by inner or outer forces. But more briefly put, this is when someone vividly feels like a natural automaton (biochemical puppet, moist robot, "meat puppet," flesh-eating zombie, etc.), or like a tool in the hands of some other powerful manipulative agent or agency. So I am saying that in order to have deeply free will, then, we must *not* have a self-consciously or reflectively fettered, epiphenomenal, will: on the contrary, we must not only *have* veridical psychological freedom, but also *be at least fully disposed to believe, or actually believe, ourselves to have an unfettered, non-epiphenomenal, real causally spontaneous will*.

Ironically, as I indicated in section 4.8, this is as true of self-styled Hard Determinists and Hard Incompatibilists as it is of everyone else. I am absolutely sure that when these philosophers choose and act, under normal conditions, they do not actually feel or believe in accordance with their own metaphysics of free will—that they do not either really feel like biochemical puppets, moist robots, "meat-puppets," flesh-eating zombies, etc., or really believe themselves to be natural automata of any kind. To be sure, the logical scope of their philosophical beliefs extends universally over all people, including themselves: but epistemically speaking, it is one thing to apply property *P* to everyone, including of course oneself, *third-personally*, and quite another thing to apply property *P* to oneself, *first-personally*. Universal instantiation is neither semantically nor epistemically equivalent to first-personal indexical predication. If *everyone* is *P* then it necessarily follows that *I* am *P*, because I am one of the people in the domain of discourse. But if everyone is *P*, even if I believe that everyone is *P*, it does not necessarily follow *that I really believe that I am P*, because I still have to identify myself with one of the many people in the domain of discourse, and I might self-consciously or non-self-consciously refuse to do that.

Otherwise put, if Hard Determinists or Hard Incompatibilists really *did* believe *that they themselves were natural automata* (biochemical puppets, moist robots, "meat puppets," flesh-eating zombies, etc.), then obviously they would seek—or at least obviously they would at least *need*—psychiatric help for the treatment of well-attested symptoms of schizophrenia, like the unfortunate Ketema Ross.[374] Furthermore, I submit that no healthy, sane real human person ever really and truly believes himself or herself, in their heart of hearts, to be a natural

automaton (biochemical puppet, moist robot, etc.), *no matter what their free will metaphysics says*.

Consider, for example, classical Compatibilism/Soft Determinism. The very idea of "real" freedom of choice or action, when taken together with a self-consciously or reflectively fettered, epiphenomenal, will, would be as absurd and pointless as freedom of action together with either a naturally determined or a powerfully manipulated will. Indeed, as Frankfurt rediscovered and influentially pointed out, psychological freedom, per se, although it is consistent with Universal Natural Determinism (namely, when it is non-veridical psychological freedom), nevertheless remains essential to our rational human personhood, as veridical psychological freedom, at the level of effective first-order desires, together with the self-conscious or reflective awareness of having veridical psychological freedom, at the level of decisive or self-identifying second-order volitions.[375] This is because a self-consciously or reflectively fettered, epiphenomenal, will utterly defeats and undermines our belief in our own intentional agency.

Now consider Hard Determinism or Hard Incompatibilism. Virtually every theory that falls under these rubrics *also* has an "debunking strategy" or "error-theory," which says that our brains mechanically create the cognitive illusion that we are really free, even though we are really natural automata. But then those theories have the following serious epistemic problem. Suppose that you hold the following view:

> Free will *is* an illusion. Our wills are simply not of our own making. Thoughts and intentions emerge from background causes of which we are unaware and over which we exert no conscious control. We do not have the freedom we think we have.[376]

If that is true, then we are all natural machines with an irresistibly strong tendency to create cognitive illusions for ourselves. Therefore, under the supposition that her theory is true, any holder of such a view *cannot rule out* the real possibility that she has created a cognitive illusion for herself by defending Natural Mechanism together with a debunking strategy or an error-theory. But if she cannot rule this out, *then she is not rationally justified in believing in her own theory*. So her belief in her own theory is cognitively self-stultifying. This conclusion, in turn, *debunks the would-be debunkers*.

It is also very important to point out in this connection that having a self-consciously or reflectively fettered, epiphenomenal, will is categorically *not* the same as the classical, Lord-Byron-style, Romantic self-consciousness of being in the grip of a grand passion that carries you away with it. The crucial contrast between these two is the inherent, categorical difference between

> (i) believing yourself to be a *mere natural machine*, that is epiphenomenally caused or helplessly manipulated by something inside your own body or outside your own body, and
> (ii) believing yourself to be *fully alive*, driven, energized, and invigorated by some naturally purposive and naturally creative vital power that is immensely bigger than you are.

Now the first-order consciousness or subjective experience of life and vitality, as I argued in chapters 2 and 3, insofar as it essentially non-conceptually and veridically picks out immanent

structural properties of non-equilibrium thermodynamic systems, is inherently anti-mechanical and uncomputable. In other words, insofar as the Lord-Byron-style Romantic phenomenology is veridical, then it entails local incompatibilism with respect to Natural Mechanism. This point is especially telling because the Lord-Byron-style Romantic phenomenology is sometimes used as an intuitively evidential ground for believing in Compatibilism and Soft Determinism,[377] including In-the-Zone Compatibilism. But this line of argument just confuses one kind of "carried away" phenomenology, with a categorically different kind of "carried away" phenomenology. Moreover, as we saw in the real-world case of Kleist, in section 4.4, this philosophical confusion can *also* be a *tragic* mistake. A natural automaton can *never* have an essentially non-conceptual and veridical Lord-Byron-style phenomenology. At best, it could only be that *his psychic motor is racing*. Indeed, Kleist's confused recognition of this point—vividly recognizing, on the one hand, the epistemic and metaphysical plight of human agents as natural automata, but on the other also mistakenly believing himself to *be* a human "meat-puppet"—tragically, drove him to suicide.

In any case, what Natural Libertarianism adds to Frankfurt's deep insight about psychological freedom are the further insights that, over and above veridical psychological freedom, per se, we *also* require the capacity for self-commitment to a live option, or Kiekegaardian Either/Or, in conjunction with our *also* being spontaneous real causes and *also* being owners of our choices or acts, in order to constitute the complex metaphysical core of our rational intentional free agency. Nevertheless, that complex metaphysical core necessarily contains veridical psychological freedom as a proper part. This in turn makes it possible to provide a fully-unpacked analysis of the self-commitment option, to the effect that my choosing or doing X includes the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, if and only if

> (i) I can choose or do X,
> (ii) I can not-choose or not-do X by just failing or refusing to choose X (= the null choice options),
> (iii) X would never actually happen (or: would not have happened) if I were not to choose (or: had I not chosen) X,
> (iv) conditions (i) through (iii) can obtain even if, in context, I have no alternative possibilities in the classical sense of branching futures and the future is temporarily not open, and finally
> (v) the phenomenology of my choosing or doing X is veridical psychological freedom.

This brings me to the third, final, and overarching level of the free agency structure, the capacity for principled authenticity, which I have already described and rationally motivated in chapter 3. This capacity includes a capacity for autonomy in the Kantian sense, namely, the innate capacity for self-legislation according to, and for the sake of, the Categorical Imperative or moral law, and for the sake of the *dignity* of real persons, namely, their absolute, intrinsic, nondenumerable, objective value.[378] Now just as psychological freedom, per se, is necessary but not individually sufficient for deep freedom, and just as deep freedom is necessary but not individually sufficient for autonomy in the Kantian sense, so too autonomy in the Kantian sense is necessary but not individually sufficient for principled authenticity. In addition to

psychological freedom, deep freedom, *and* Kantian autonomy, my choosing or doing *X* must also have two further features.

First, I must have the will that I want, or what Frankfurt calls a *decisive identification* between my second-order volitions and my first-order effective desires.[379] This necessarily includes my having a true first-person-indexically-self-predicating, occurrently self-conscious, belief about my own first-order veridical psychological freedom. It is also the same as what Existentialists have called "purity of heart" or *authenticity*, and what Frankfurt himself calls "wholeheartedness,"[380] when we take into account the dynamic extension of decisive identification over the temporal duration that is "the time of their lives"—because it essentially involves someone's living a life of passionate, self-realizing, single-minded adherence to her own principles, together with her taking complete deep (non-)moral responsibility for some brute facts over which she had no control.[381]

Second, my choosing or doing must also have what Kant calls *moral worth* because I actually choose or do *X* essentially from respect for the dignity of real persons and for the moral law, over and above whatever other desires I may normally be moved by. Respect, in turn, as we have seen in sections 3.3 and 3.4, is an innate universal emotional disposition in rational human animals to generate consciously-experienced second-order volitions to be moved to choose and do things by non-egoistic, non-hedonic, and non-consequentialist, morally good and right first-order desires, in place of egoistic, hedonic, or consequentialist, first-order desires, especially if they are morally bad and wrong, that would have moved us instead.[382] It is extremely important to remember here, in view of Kantian non-intellectualism, that egoistic, hedonic, or consequentialist first-order desires *are not, in and of themselves, morally bad or wrong*. They can, indeed, in a given context, move us to choice or action that is, in that context, for the sake of the moral law; and they could also, in a given context, move us to choice or action that is, in that context, merely in conformity with the moral law and not for the sake of the moral law. But they can *also*, as a matter of "radical evil" and the "human, all too human" perversity of the heart and will, move us to choice or action that is morally bad and wrong, banally evil, or even near-satanically evil. Hence the crucial point about the moral worth of choice or action, and being motivated by respect, is that respect *really can constitute* an essentially non-conceptual and life-changing "revolution of the heart" or "revolution of the will," that, in turn, is triggered by our self-conscious or reflective, conceptual recognition of the Categorical Imperative as a desire-overriding, strictly universal, a priori, categorically normative non-instrumental reason for action. It is a sad, true fact about the rational human condition that we rarely do this; but it is an equally sublime true fact about the rational human condition that we *do* do this much more often than you might think.[383]

Sad and sublime realities aside, it nevertheless remains fully true that satisfaction of these two conditions, together with all the other necessary conditions of deep freedom, jointly constitute what I call "principled authenticity." This is because, when taken all together, they jointly constitute not only a wholehearted adherence to my own principles, but also a wholehearted adherence to some *absolutely universal objective* moral principles—for example, the several formulations of the Categorical Imperative—along with my taking complete responsibility for some brute facts over which I had no control.

As I noted in chapter 3, by a categorical contrast, the moral contrary of authenticity in the sense of a wholehearted adherence to principles, namely, *inauthenticity*, is comporting yourself as if you were a natural automaton—as if you were a human turnspit, or a fleshy deterministic or indeterministic Turing machine, and not really alive and caring; as if you could never think

or choose or act for yourself; and as if you did not really have the capacity for deep freedom, deep (non-)moral responsibility, and principled authenticity. In the *Philosophical Investigations*, Wittgenstein asks, "Couldn't I imagine having frightful pains and turning to stone while they lasted?"[384] Correspondingly, to get a sense of the nature of inauthenticity, now imagine yourself *having frightfully strong desires for something*, whether this thing has high egoistic, hedonic, or consequentialist value, or is instead specifically a target of moral respect, and then *turning into a biochemical puppet and moist robot* "while they lasted." Indeed, it is precisely this thought-experiment that it is stunningly artistically expressed by the original version of the 1950s science-fiction classic, Don Siegel's *Invasion of the Body-Snatchers.*[385]

In view of what I said, just above, about the equally sad and sublime true facts about us, to say that principled authenticity is "really possible" for us, however, is *not* to say that we are *always* or even *usually* moved by respect for human dignity and by higher-order love for the moral law that is inscribed in our rational but still "human, all too human" hearts. Usually we are moved only by egoism, hedonism, or purely instrumental concerns. But *by no means are we always* moved by egoism, hedonism, or purely instrumental concerns. Hence as rational human animals or real human persons with deep freedom of the will (ultimate sourcehood, up-to-me-ness), and with an innate capacity for principled authenticity, necessarily we *really can* be so moved, because we *ought* to be so moved, and, under certain actual contextual and historical conditions, we *are* so moved. It is quite true that "from the crooked timber of humanity, nothing straight can ever be made" (*IUH* 8:23); but it is equally true *that crooked timber can sometimes be sufficiently strong timber for the purposes at hand*. In other words, all rational human animals, or real human persons, are "human, all too human," "radically evil," and "miserable sinners." But this is not only fully consistent with, but also necessarily complementary with—the flip side of—our possessing the capacity for achieving principled authenticity at least partially or to some degree, and, in propitious circumstances, for achieving a sublime "sinner-sainthood."[386]

In any case, as everyone knows, *ought* does not entail *is*. But on a Kantian account of the relationship between morality and deep freedom/deep moral responsibility, or at least on the contemporary Kantian account of practical agency, especially including The 2D Conception of Rational Normativity, that I have spelled out in chapter 3, *ought* entails a real *can*; and empirical, historical evidence about actual humanity confirms its objective reality. In this way, deep freedom (including both veridical psychological freedom and also the capacity for self-commitment to a live option, or Kierkegaardian Either/Or) and the capacity for principled authenticity, are jointly necessary and sufficient conditions of our deep (non-)moral responsibility, and jointly necessary but not jointly sufficient conditions of the moral goodness and rightness of our choices and acts. Nevertheless the advance to sufficiency actually happens when, in propitious circumstances, we really and truly do wholeheartedly choose or act from respect and for the sake of the moral law, at least partially and to some degree, and achieve sublime sinner-sainthood. *Ought* entails *can*, and, on occasion, *we really do do what ought to be done*.

5.4 CONCLUSION

This completes my six-step overall argument for the truth of Natural Libertarianism, which, as per section 1.3 above, looks like this:

> The Six-Step Argument for Natural Libertarianism
>
> (1) Beyond Mechanism.
>
> Biological life is a physically irreducible but also non-dualist and non-supervenient necessary a priori immanent structure of a well-defined class of far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing thermodynamic systems. (Premise, justified in chapter 2.)
>
> (2) From Biology to Agency.
>
> Free rational minded animal agents are nothing more and nothing less than conscious, intentional, caring, rational self-organizing, organismic thermodynamic systems that are capable of (i) deeply free choice based on effective desires and instrumental or non-instrumental internal reasons, (ii) autonomy in the Kantian sense, or rational self-legislation, and (iii) authenticity in the Existentialist sense, that is, purity of heart, single-mindedness, or wholeheartedness. (Premise, justified in chapter 3.)
>
> (3) Neither/Nor.
>
> Natural Mechanism is the weak disjunctive combination of Universal Natural Determinism and Universal Natural Indeterminism. More specifically, something is naturally mechanized, or a natural automaton, if and only if all its causal behaviors, functions, and operations are necessarily determined by all the deterministic or probabilistic/statistical general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and-or-energy facts about the past, especially including The Big Bang, and Turing-computable from that base. But not everything natural is Conservation-Laws-determined, Big Bang-caused, and Turing-computable. So Natural Mechanism is false, hence both Universal Natural Determinism and Universal Natural Indeterminism are false. Morever, Hard Determinism is false. Soft Determinism is false. And Classical Libertarianism (inlcuding its agent-causal, non-causal, and event-causal indeterminist versions) is false. Correspondingly, classical Compatibilism (including Soft Determinism, Semi-Compatibilism, Revisionism, and In-the-Zone Compatibilism) and classical Incompatibilism (including Hard Determinism, Hard Incompatibilism, classical Agent-Causal Libertarianism, Non-Causal Indeterminism, and Event-Causal Indeterminism) are all false. At the same time, Local Incompatibilism and Non-Local Compatibilism are true. So Incompatibilistic Compatibilism is true. (Premise, justified in chapter 4.)

(4) Either/Or.

Harry Frankfurt's argument against The Principle of Alternative Possibilities (PAP) is sound, but it does not follow that deep (non-)moral responsibility is compatible with either Universal Natural Determinism or Universal Natural Indeterminism, since Natural Mechanism is false. On the contrary, the Kierkegaardian Either/Or, which flows from the capacity for self-commitment to a live option, is presupposed by all Frankfurt-style counterexamples to PAP; and as metaphysically embedded in a larger free-agency-structure which also includes the capacities for veridical psychological freedom and for principled authenticity (that is, the capacity for autonomy in the Kantian sense, or rational self-legislation, together with the capacity for purity of heart, single-mindedness, or wholeheartedness), this capacity for self-commitment to a live option, or Kierkegaardian Either/Or, is a necessary and sufficient condition of deep (non-) moral responsibility. (Premise, justified in chapter 5.)

(5) Deep Freedom and Principled Authenticity.

The capacity for self-commitment to a live option, or the Kierkegaardian Either/Or, along with the capacities for veridical psychological freedom, real causal spontaneity, and ownership, are necessary and sufficient conditions of the capacity for deep freedom. In turn, the capacity for deep freedom is a necessary but not sufficient condition of the capacity for principled authenticity, which, as incorporating deep freedom, yields deep (non-)moral responsibility (Premise, justified in chapter 5.)

(6) Natural Libertarianism.

Therefore, since Natural Libertarianism is just the three-part thesis (i) that freedom is in life, (ii) that Incompatibilistic Compatibilism is true, and (iii) that the constitution of free rational human minded animal agency inherently includes the capacities for deep freedom and principled authenticity, together yielding deep (non-)moral responsibility, then it follows that Natural Libertarianism is true. (Conclusion, from premises 1-5 above.)

It should be evident by now that Natural Libertarianism is sharply distinct from each of The Three Standard Options, as well as from Semi-Compatibilism, In-the-Zone Compatibilism, Hard Incompatibilism, and Revisionism. At the same time, Natural Libertarianism incorporates some non-trivial aspects of classical Compatibilism in its non-classical non-local compatibilism, and also incorporates some non-trivial aspects of classical Incompatibilism in its non-classical local incompatibilism. Natural Libertarianism also incorporates some non-trivial aspects of Soft Determinism by

(i) requiring *an anti-luck condition* on free choosing and doing—a condition that is secured by the conjunction of anti-mechanism and the self-commitment to a live option, or Kierkegaardian Either/Or—and

(ii) zeroing in on free agency in the *actual* event-sequence and ignoring *counterfactual* sequences.

And Natural Libertarianism *also* also incorporates some non-trivial aspects of Classical Libertarianism in its non-deterministic conception of free agency and its metaphysical appeal to real causal spontaneity. So Natural Libertarianism, in effect, preserves whatever there was in the classical and standard views that is actually true and worth preserving.

But perhaps most importantly of all, Natural Libertarianism provides a metaphysically revisionary, liberally naturalistic, naturally pietistic, and altogether *philosophically emancipatory and "woke"* way of thinking and feeling about free agency. According to this radically enlightened way of thinking and feeling, free agency is a physically irreducible and anti-mechanistic, but also non-dualistic and non-supervenient, immanent structural fact about a special, well-defined class of non-equilibrium thermodynamic systems: the organismic, finegrainedly normatively attuned, minded, 2D rational, deeply free, deeply (non-)morally responsible ones—namely, *real persons who are capable of achieving principled authenticity, at least to some salient degree or extent*. So our free agency is nothing more and nothing less than our freedom-in-life. This, in turn, means that the natural world, even despite its being thoroughly nonideal, is not in any way epistemically or metaphysically alien to us. On the contrary, the natural world is *our town*.

By a two-part sharp contrast, on the one hand, ontological dualism epistemically and metaphysically alienates us from nature in the macrocosmic, outer sense, the larger physical world, hence it alienates us from "the starry heavens above me" (the rock); and on the other hand, physicalism turns us into natural automata and epistemically and metaphysically alienates us from nature in the microcosmic, inner sense, our own embodied agentive selves, hence it alienates us from "the moral law within me" (the hard place). Natural Libertarianism therefore fully *avoids* both the outer-alienating rock and the inner-alienating hard place.

Natural Libertarianism's biologically-grounded, anti-mechanistic, non-reductive, non-dualist, non-supervenience-based account of free agency also effectively incorporates a rich conception of practical agency, according to which more-or-less wholehearted human caring is the essentially embodied vital engine of pure practical reason, under the active guidance of self-conscious or reflective, deliberative practical reasoning that is inherently governed by absolutely universal, a priori non-instrumental moral principles.

Furthermore and finally, this philosophically emancipatory and woke way of thinking about free agency shows us that the deeply-ingrained classical Compatibilism vs. classical Incompatibilism dichotomy was a scandalously false *di*chotomy, and that the Hard Determinism vs. Soft Determinism vs. Classical Libertarianism trichotomy was also a scandalously false *tri*chotomy, while at the same time substantively connecting itself with both the Kantian and Existentialist traditions alike.

And all of this, in turn, guarantees the inseparable fusion of the real metaphysics of free agency with the real metaphysics of what I call *real personhood*, which I now want to investigate explicitly and in detail.

Chapter 6

MINDED ANIMALISM I: WHAT REAL PERSONS REALLY ARE

> Beings the existence of which rests not on our will but on nature, if they are beings without reason, still have only a relative worth, as means, and are therefore called *things*, whereas rational beings are called *persons*, because their nature already marks them out as an end in itself, that is, something that may not be used merely as a means, and hence so far limits choice (and is an object of respect)…. If, then, there is to be a supreme practical principle, and, with respect to the human will, a categorical imperative, it must be one such that, from the representation of what is necessarily an end for everyone because it is an *end in itself*, it constitutes an *objective* principle of the will and thus can serve as a universal practical law. The ground of this principle is: *rational nature exists as an end in itself.* (*GMM* 4: 428-429)

> It is my view that one essential difference between persons and other creatures is to be found in the structure of a person's will.[387]

> The Biological Approach is the view that you and I are human animals, and that no sort of psychological continuity is either necessary or sufficient for a human animal to persist through time.[388]

6.0 INTRODUCTION

In this chapter and the final one, I want to work out the real metaphysics of specifically *real* persons—as opposed to logically, conceivably, analytically, or "weakly metaphysically" possible persons that are non-animals, disembodied, or even divine, on the one hand, and also as opposed, on the other hand, to

> either (i) actual *artificial* persons, created by human convention, like public personae (for example, Cary Grant, or Mark Twain) and public offices (for example, The President of the USA, or The Prime Minister of Canada),[389]
> or (ii) actual *collective* persons, also created by human convention, such as legal bodies (for example, The Supreme Court of the USA, or The European Court of Human Rights), governments (for example, The US Senate, or The British House

> of Commons), and business corporations (for example, Amazon, Apple, Google, Microsoft, etc., etc.).

And I also want to work out this real metaphysics with special reference to real *human* persons, that is, with special reference to *us*, but also with a full acknowledgment of the fact that not necessarily all real persons are human.

(As before, from here on in, unless otherwise specified, and apart from a few places where I use the phrase "real metaphysics" for special emphasis, by using the term "metaphysics" I always mean *real metaphysics* in the contemporary Kantian manifest realist, anti-noumenal-realism, anti-Standard Picture, anti-scientific naturalism/X-Phi/second philosophy sense, that I spelled out in section 1.0.)

In order to undertake the metaphysics of real persons, however, it is crucial right from the start to distinguish carefully between

> (i) the metaphysics of real *personhood*,[390] and
> (ii) the metaphysics of real *personal identity*.[391]

The metaphysics of real personhood corresponds to the question, "what is the nature of a real person?" I will call this *The "What-am-I?" Question*. Once an answer to The "What-am-I?" Question has been determined, by means of identifying a set of necessary and sufficient conditions for membership in the relevant real personhood class or kind, then we can determine the class or kind of all actual and possible real persons. The metaphysical topic here is *the essence of a kind*. According to my view, in the long-winded version, real persons essentially are

> conscious, intentional, caring, 2D rational or sapient, far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, finegrainedly normatively attuned thermodynamic systems, with non-mechanical, uncomputable, physically irreducible, yet also non-dualist and non-supervenient properties, whose choices and acts are inherently constrained, guided, and governed by self-legislated categorically normative logical and moral principles, and whose highest aim is the achievement of principled authenticity, at least partially or to some degree.

Thus according to my view, but now in the short-winded version, real persons essentially are *rational minded animals with freedom-in-life*. This metaphysics of real personhood constitutes the first part of the overall doctrine about real persons that I call *Minded Animalism*.

The metaphysics of real personal identity, by contrast, corresponds to the importantly distinct question, "which *one* of the real persons am I?" I will call this *The "Who-am-I?" Question*. Here we are looking for a necessary and sufficient criterion that singles out one and only one actual member of the class or kind of all actual and possible real persons, and determines her persistence over time. The metaphysical topic here is individuation under a kind, together with that individual's temporal persistence. According to my view, a real person like you or me is identical *to each and every proper part of her own complete, finite, and unique rational human life*, and correspondingly also identical *to her whole complete, finite, and unique rational human life*.[392] This life begins when a living organism belonging to a certain

species—in our case, of course, the human species—becomes conscious or minded, continues through the eventual manifestation or realization of her capacities for minded animal free volition, and then rational animal free agency, and then ends with her permanent death.[393] In section 7.2 below, I will characterize this fundamental fact about us more precisely and rigorously by arguing that our real human personal identity is a sub-kind of the mereological relation of *metonymous identity* (that is, part-to-whole identity) between proper (spatio)temporal parts and their corresponding unified (spatio)temporal wholes. But in the meantime, it is sufficient to say that I am claiming that *we are identical to each and all of the stages of our own complete, finite, and unique rational human animal lives*.

My view on the metaphysics of real personal identity is a significant extension and strengthening of what Eric Olson calls The Biological Approach, or "Animalism," to all and only rational minded animals, that is, conscious, intentional, caring, rational living organisms, whose mental lives are necessarily and completely neurobiologically embodied, and who are strictly identified with each and all of the stages of their minded animal lives. Standard Animalism identifies us with individual living human animals that only contingently have mental properties, including consciousness and perhaps also rationality, and that also possess these properties only during certain phases of their animal lives. By contrast, according to the view I am calling *Minded* Animalism, real human persons are identified with each and all of the stages of the lives of individual living conscious, intentional, caring, rational human animals *who, by their very nature, necessarily have conscious minds that are essentially embodied*.[394] This substantive metaphysics of real personal identity, including its significant extension and strengthening of standard Animalism, constitutes the second part of Minded Animalism.

The basic distinction between the metaphysics of real personhood and the metaphysics of real personal identity also leads on directly to three other important preliminary points.

First, a correct answer to The "What-am-I?" (namely, real personhood) Question does not in and of itself yield a correct—or indeed any—answer to The "Who-am-I?" (namely, real personal identity) Question. The term "person" is a kind term or sortal term that is also normally associated with identity conditions for individual persons falling under that kind. But satisfying the conditions for belonging to a given kind (for example, water) does not in and of itself guarantee the satisfaction of identity conditions for individuals (for example, drops of water, lakes, rivers, or oceans) falling under that kind. So there is no necessary entailment from the metaphysics of real personhood to the metaphysics of real personal identity.

Second, a correct answer to The "Who-am-I?" (namely, real personal identity) Question presupposes a correct answer to The "What-am-I?" (namely, real personhood) Question, but not (as we just saw) conversely. As a matter of real metaphysics, it is presuppositionally necessary to determine the class or kind of all actual and possible real persons, prior to determining which individual one of them I am, and how I persist over time. But, again as a matter of real metaphysics, it is not presuppositionally necessary to determine just which individual one I am, within the class or kind of all actual and possible real persons, and how I persist over time, in order to determine the class or kind of all actual and possible real persons. Therefore, in order to give a correct real metaphysics of real personal identity, one must first provide a correct real metaphysics of real personhood. Trying to go the other way, either by inferring immediately from prima facie rational intuitions about personal identity to a metaphysical theory of real personhood, or by giving a free-floating metaphysics of real personal identity without also considering its necessary grounding in a real metaphysics of real personhood, merely begs the question.

It should also be stressed that by "prima facie rational intuitions about personal identity," I am *not* talking about our ordinary, pre-reflectively conscious, deeply-held sense that we are real persons and can be strictly identified as persisting over time, by reference to the proper parts or the whole of our unique, finite, and essentially embodied *lives*.[395] This ordinary pre-reflectively conscious sense of ourselves as being identified *with our lives* is every bit as deeply-held as the sense that we are free agents. Indeed, both senses are phenomenologically self-evident. Furthermore, the two senses flow into one another in a complementary way, and both belong to real metaphysics.

Third and finally, whereas the real metaphysics of real personhood is the real metaphysics of a kind or sort of entity, according to my view, the real metaphysics of real personal identity is the real metaphysics of each and all of the stages of the whole life-process of some individual member of that kind. More precisely, according to my view the real metaphysics of real personhood is the real metaphysics of *a kind-constituting structure*, whereas the real metaphysics of real personal identity is the real metaphysics of *a necessarily diachronic entity, namely, a biophenomenologically dynamic and thermodynamic process*—bearing the real personhood structure, to be sure, but not the same as the structure itself—that is inherently spread out and forward-directed in actual space and actual time. Real personal identity, as I am understanding it, necessarily comprehends an egocentrically-centered, essentially embodied, *phenomenologically dynamic* life-process of a conscious, intentional, caring, 2D rational living organism that is inherently living in an intrinsically directional or orientable space, while also "having the time of its life," in such a way that this unified far-from-equilibrium, complex, self-organizing *thermodynamic* life-process necessarily also has temporal irreversibility. In other and fewer words, our real human personal identity, just like our minds and our free agency, is nothing more and nothing less than a form of life that grows naturally in our animal bodies. Or in still other and even fewer words, *I am my life, for better or worse*.

On this account, every real person's life begins in a unique birth and ends in a unique death. That is, the life of every real person has a definite spatiotemporal beginning or birthplace and time of birth, and also a definite ending or death-place and time of death. Correspondingly, the real metaphysics of real personal identity is the real metaphysics of a special kind of real-personhood-structured spatially orientable and temporally asymmetric finite far-from-equilibrium, complex, self-organizing, finegrainedly normatively attuned, minded animal thermodynamic life-process, and not merely the real metaphysics of the kind-constituting real personhood structure alone. Hence I am nothing more and nothing less than the egocentrically-centered, spatially orientable, forward-directed Little Bang that is each and all of the stages of my own complete, finite, and unique rational human animal life. Each one of us is thereby *special enough* for all metaphysical, epistemological, normative, and moral purposes—but not so very, very special, after all. In particular, we are not ghostly souls, floating above the physical world. But neither are we essentially like the chair I am sitting on, or the laptop computer I am typing on. We are neither "something over and above the physical" (ghosts) nor "merely physical" (hunks of lifeless mechanical matter, or inert material, temporally-reversible, equilibrium thermodynamic processes, namely, natural automata). We are neither ghosts nor machines! Instead, we are living organisms, which means that we are fully natural and physical, yet "not so damned physical," in the way that natural automata are, and more specifically, that we are *rational human minded* living organisms.

And the same same point generalizes smoothly beyond all rational human minded animals, to all rational minded animals or real persons whatsoever, of any species. It is deeply

metaphysically important neither to *inflate* or overestimate ourselves, for example, via ontological dualism or speciesism, nor to *deflate* or underestimate ourselves, for example, via reductive or non-reductive physicalism and natural mechanism.[396] That way—forever suspended between the outer-alienating rock of ontological dualism/specieism (the ghost) and the inner-alienating hard place of physicalism/natural mechanism (the machine)—metaphysical madness lies. To recover and to preserve our metaphysical sanity, we must be able to see ourselves *as we really are*, namely, as *rational minded "human, all too human" life-forms*.

6.1 FROM DEEP FREEDOM TO REAL PERSONS

According to the real metaphysics of rational human animals that I am proposing, the fundamental concept and fact of *free agency* (= free will + practical agency) and the fundamental concept and fact of *real personhood*, as mediated by the fundamental concept and fact of *real personal identity*, are all bound together at their very cores, and entail each other with non-logical, essentially non-conceptual, synthetic a priori, "strong metaphysical" necessity. So the internal structure of that metaphysics looks like this diagram, when read either downwards from the top line to the bottom line, or upwards from bottom line to the top line:

FREE AGENCY
synthetically a priori necessarily if and only if
REAL PERSONAL IDENTITY
synthetically a priori necessarily if and only if
REAL PERSONHOOD

Free agency, especially including deep freedom and deep (non-)moral responsibility, together with the capacity for principled authenticity, was the special topic of chapters 1 to 5, and it yielded the doctrine of Natural Libertarianism. In this chapter, and in the next and final chapter, I want to work out a unified theory of free agency and real personhood, by way of the mediating concept and irreducible, primitive fact of real personal identity. The four leading ideas of this theory are as follows:

(*Leading Idea 1*) Real persons are absolutely, intrinsically, nondenumerably, objectively valuable, that is, real persons have *dignity*
(*Leading Idea 2*) A real person is a conscious, intentional, caring, 2D rational animal capable of free agency, and there are two distinct types of real persons, namely, (2a) *non-autonomous*, "lower-level," or *Frankfurtian* real persons, aka real persons$_f$, and (2b) *autonomous*, "higher-level," or *Kantian* real persons, aka real persons$_k$.
(*Leading Idea 3*) The three individually necessary, individually insufficient, and jointly sufficient conditions of the identity of a real person are (3a) the intrinsically spatiotemporal continuity of that 2D rational animal's consciousness, intentionality, and caring (*biophenomenological continuity*), together with (3b) the intrinsically spatiotemporal continuity of that 2D rational animal's necessary and complete, hence essential, embodiment (*biological and neurobiological continuity*), together with (3c) the further

> fact that *enough* of the dual biophenomenological and biological/neurobiological events constituting the continuous life of that essentially embodied 2D rational animal are also deeply free, involving ultimate sourcehood, and up to her (*freedom-dominance*).
> (*Leading Idea 4*) Real personal identity is a specification of the mereological relation of *metonymous identity* between proper (spatio)temporal parts and unified (spatio)temporal wholes, such that a real person is identical to each and all of the stages of her complete, finite, and unique 2D rational animal life.

Srictly speaking, in the human species at least, there is also a class of *semi-autonomous* or "middle level" real persons, falling between the class of non-autonomous, "lower-level," or Frankfurtian real persons (persons$_f$) and the class of autonomous, "higher-level," or Kantian persons (persons$_k$). This is the class of *adolescents* or *teenagers*. But this class, at bottom, is really a set of borderline cases between the two fundamental classes of real persons, and does not collect an essentially *different* kind of real persons. See section 6.3 below for more details.

As to the first leading idea, it is obviously closely historically related to Kant's ethics; and it also is a central feature of my version of contemporary Kantian ethics, as worked out in *Kantian Ethics and Human Existence*.

Equally obviously, the second leading idea, in turn, has important similarities with Frankfurt's hierarchical-desire theory of personhood—with one crucial difference, which is that on my account, as I mentioned just above, there are two importantly different classes of real persons, namely,

> (i) non-autonomous, lower-level, or Frankfurtian real persons, real persons$_f$, and
> (ii) autonomous, higher-level, or Kantian real persons, real persons$_k$.

The class of lower-level or Frankfurtian real persons is paradigmatically exemplified by most human children, but also by some non-human animals, for example, Great apes (by which, again, I mean non-human members of the biological family Hominidae, including bonobos, chimpanzees, gorillas, and orangutans), and perhaps also dolphins. Lower-level or Frankfurtian real persons are, as it were, *junior* real persons; and in relation to them, the higher-level or Kantian real persons, paradigmatically exemplified by most human adults, are as it were *senior* persons. Similarly, there can be junior and senior members of the same family, club, team, academic department, college, university, law firm, or other business corporation.

I do not want to push this analogy *too* hard, however, because of the various kinds of metaphysically irrelevant and sometimes also quite invidious, rationally-unjustified, and immoral social-status and social-power values assigned to roles within families, clubs, teams, academic departments, universities, etc. The crucial point I want to make is just that the two types of persons have *ranked equality* of moral status. They are both equally real persons, and correspondingly they are both equal in absolute, nondenumerably infinite, intrinsic, objective moral value, or dignity. This means that, like transfinite cardinal quantities, the value of each real person cannot be compared with, made equivalent to, or exceeded by, any denumerable economic value quantity. Nevertheless, even though no real person is more or less valuable than any other real person, and even though a real person has no economic value equivalent, or price, again like the series of transfinite cardinals, *more* real persons are still more valuable than

fewer real persons. So, other things being equal, it is morally better to help more people than it is to help fewer people, and it is also morally worse to harm more people than it is to harm fewer people. In any case, as possessors of dignity, all real persons must be *treated with respect*, and morally *considered* equally, although not always *treated* equally. What respect for human dignity requires by way of treatment, is that, other things being equal, *people never be treated like mere means or mere things, and also that they always be provided with what is sufficient for their true human needs*. But people have different *specific* needs at different times and in different places, even while having the same *generic* true human needs. Hence respect for human dignity, and moral equality, are importantly different.[397] Moreover, the basic capacities of real persons$_f$ and real persons$_k$ are, respectively, somewhat differently configured and disposed; and correspondingly their moral duties, moral responsibilities, their moral consideration and treatment of each other, and their ideal moral aims or goals, both with respect to themselves and to one another, are also somewhat differently specified.

More specifically, only autonomous, higher-level, or Kantian real persons are capable of the kind of mutually recognizing, self-conscious respect, and deep (non-)moral responsibility for their choices and acts, that is characteristic of any degree of principled authenticity. Infants, other young children, and real persons that are non-human animals, are simply not capable of this. It would be madness, for example, to expect any five year-old child, or a chimpanzee, to be able to take on the exceptionally demanding moral-political role of the President of the USA.[398]

The third leading idea, as I mentioned above, is a significant extension of Olson's Biological Approach to the metaphysics of personal identity, or what I will call *standard Animalism*. Standard Animalism says that biological continuity is necessary and sufficient for the identity of persons, who contingently have mental properties.[399] Or otherwise put, standard Animalism says that I am literally identical with an individual human living organism, my human animal, which, for some portion of its life, is also capable of thinking. My view, Minded Animalism, shares with standard Animalism the thesis that biological continuity, and thus continued individual animal life, is a necessary condition of personal identity. So I am literally identical with (each and all of the parts of the life-process of) a creature that is also an individual human animal. But virtually all contemporary versions of standard Animalism are also committed to reductive or non-reductive materialism about conscious minds and persons, whereas Minded Animalism, in view of its basic commitment to the essential embodiment theory of the mind-body relation and its neo-Aristotelian hylomorphism, rejects both of these. Moreover, by a significant extension of standard Animalism, according to Minded Animalism, although biophenomenological continuity is indeed necessary, it is also *individually insufficient*, and only when it is taken together with essentially embodied biological/neurobiological continuity and also together with something I call *freedom-dominance*, are these three factors jointly sufficient for personal identity.

Finally, the fourth leading idea is a formal interpretation of the classical Existentialist conception of a real person's life, as applied to the metaphysics of personal identity. In Sartre's apt words, this says that

> [the human being] is nothing other than his plan; he exists only to the extent that he fulfills himself; he is therefore nothing other than the ensemble of his acts, nothing other than his life.[400]

According to the docrtine of a specifically *Minded* Animalism, as opposed to standard Animalism, I am not literally identical with any individual living human animal in and of itself, in the sense of that animal's either being taken apart from its having mental properties; or being taken apart from each and all the stages of its essentially embodied conscious, intentional, caring, 2D rational life-process; and or being taken apart from its free agency. Those separations are non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" impossible. Thus I am not identical with the non-minded first or second trimester human fetus that later became me. I am not identical with a non-minded individual human animal that lives on after my upper brain has been destroyed in an accident or removed by surgery, despite the fact that this creature conventionally bears my proper name. I am not identical with any minded individual human animal that has had its mental life wiped clean by some brain-washing process, and then replaced by someone else's mental life. I am not identical with my corpse, again despite the fact that it conventionally bears my proper name—even if it were resurrected as a flesh-eating zombie in some future George Romero-imagined "night of the living dead." And above all, I am not identical with any physical counterpart of me that is really a natural automaton.

Perhaps most importantly, Minded Animalism is designed to close the gap opened up by Locke's famous distinction between "the identity of the person" and "the identity of the human being." Locke's distinction leads to a seemingly unresolvable dichotomy between the classical *Psychological Approach* to personal identity, according to which some psychological relation is necessary and sufficient for personal identity, and what Olson calls the classical *Somatic Approach*, according to which, on the contrary, no psychological relation is necessary or sufficient for personal identity and some fundamentally physical relation is necessary and sufficient for personal identity.

On the one hand, the basic worry about the classical Psychological Approach is that it allows for the following three problematic possibilities:

> (i) many different persons can simultaneously or successively occupy one body (for example, multiple personality disorder cases, Locke's Prince and Pauper case, split brains, etc.),
> (ii) one person can simultaneously or successively occupy many bodies (for example, Parfit's Transporter cases, fission cases or split brain transplants, etc.), and
> (iii) there can be a discontinuity of persons over time (for example, disappearance of the person during unconsciousness, amnesia, temporary insanity, etc.).

On the other hand, however, the basic worry about the classical Somatic Approach, including standard Animalism, is that it allows for the following three sharply different, but equally problematic, possibilities:

> (i) there must be a strict identity between the human person and a non-minded human fetus in the first and second trimesters,
> (ii) there must be a strict identity between the human person and an individual human animal that lives on after its higher brain has been destroyed in an accident or has been removed by surgery, or that has had its mental life has been wiped

> clean by some brain-washing process and replaced by someone else's mental life, and
> (iii) there must be a strict identity between the human person and either its corpse or its resurrected flesh-eating zombie, or any physical counterpart of it that is also really a natural automaton.

Here are two quick follow-up comments on these problematic possibilities.

First, when I say it is "problematic" that according to standard Animalism there must be a strict identity between the human person and a non-minded human fetus in the first and second trimesters, I do not, in this philosophical context, mean that this is *morally* problematic because of its implications for the morality of abortion and infanticide.[401] I mean merely that it is *metaphysically* problematic to hold that I, a real person who is inherently a *minded* animal, am strictly identical with a *non-minded* animal.

Second, strictly speaking, standard Animalism does not *officially* require that I be identical with my corpse or resurrected flesh-eating zombie, and as a consequence, standard Animalists have made serious attempts to solve The Corpse Problem.[402] But even if standard Animalism can solve The Corpse Problem, it (namely, standard Animalism) still does quite implausibly require that I be strictly identical with my unthinking first and second trimester fetus, strictly identical with the individual human creature that lives on after my upper brain has been destroyed by an accident or removed by surgery, and also strictly identical with my naturally mechanized physical counterpart. So in addition to solving The Corpse Problem, standard Animalism also needs to be able to give a non-ad hoc explanation of why personal identity *fails* to survive the difference between

> (ia) being a living material composite, and
> (ib) being a non-living material composite,

when at the same time by a standard Animalist hypothesis it *does* survive what seems to be the equally important difference between

> (iia) being a minded material composite, and
> (iib) being a non-minded material composite,

and also when at the same time by a standard Animalist hypothesis it again *does* survive the equally important difference between

> (iiia) being a natural automaton, a biochemical puppet or moist robot, and
> (iiib) being a real living animal.

Finally, here is another serious problem for standard Animalism. As of 2018, new biomedical evidence suggests that all women who are capable of becoming pregnant are in fact *totipotent* and *chimeras*, in that their DNA changes when they become pregnant, fusing with the DNA of the zygote and fetus, so that their biological individuality is not fixed until they have become

> either (i) pregnant

> or (ii) incapable of becoming pregnant.[403]

If that is correct, then many or even most women do not have a unique living animal body until several decades after they are already real persons. This is a serious problem for Standard Animalism, which identifies people with individual living animal bodies, since it would then follow that many or even most women are *not people* for much of their lives—which is clearly absurd. But it is not a problem for Minded Animalism, which identifies people with each and all proper stages of their minded animal lives.

In any case, now having very briefly surveyed the negative arguments against the Psychological Approach, the Somatic Approach, and standard Animalism, I want to present a positive argument for Minded Animalism. In his insightful paper, "Animalism," Andrew Bailey offers an *evidential* argument—that is, an argument for *believing* some proposition *P*, as opposed to a direct argument for the truth of *P* per se—for (more or less) standard Animalism, from what he calls the "association" of myself with my animal body, in response to various criticisms of the standard Animalist doctrine.[404] I do fully agree with Bailey that his pro-Animalist argument-strategy is generally effective against those criticisms. I also think that he provides some good prima facie positive reasons for believing (more or less) Standard Animalism. And I also fully agree with him that although either reductive or non-reductive physicalism is a shared assumption of virtually all contemporary versions of Animalism, nevertheless, when we historically consider classical versions of Animalism—in, for example, the Aristotelian hylomorphic tradition—this clearly shows that *Animalists need not be committed to any version of physicalism, whether reductive or non-reductive*.

But as *against* Bailey's account, according to Minded Animalism, the relation between real human persons and their essentially embodied minded animal lives is profoundly more intimate than mere "association," or Humean constant conjunction. So, following up on that critical thought, here is an evidential argument for Minded Animalism that I call *The Ecce Homo Argument*. This argument is so-named for two reasons:

> (i) Pontius Pilate's scornful words, "*ecce homo*!," usually translated as "behold the man!," as he presents a bound, scourged Jesus Christ, crowned with thorns, to angry mob, just prior to the Crucifixion, as reported in the Bible at John 19:5, and (ii) Nietzsche's classic Existentialist text, *Ecce Homo: How One Becomes What One Is*.

In other words, the evidential content I am drawing upon has significant religious, Existential, and literary echoes. More specifically, however, the argument is two-step. The single premise is a conjunction of sentences derived from Shylock's amazing speech in Shakespeare's *The Merchant of Venice*, act 3, scene 1, lines 58-68, by substituting "human minded animal" for "Jew," "real human person" for "Christian," and "be identical to" for "resemble." In turn, the inferential strategy I am using is a direct appeal to phenomenological self-evidence. And finally, the conclusion is intended to express an authoritative rational intuitive justification for belief.[405] So, projected onto that backdrop, here is the argument.

The Ecce Homo Argument for Minded Animalism

> (1) I am [a human minded animal]. Hath not [a human minded animal] eyes? Hath not [a human minded animal] hands, organs, dimensions, senses, affections, passions? Fed with the same food, hurt with the same weapons, subject to the same diseases, healed by the same means, warmed and cooled by the same winter and summer as a [real human person] is? If you prick us, do we not bleed? If you tickle us, do we not laugh? If you poison us, do we not die? And if you wrong us, do we not revenge? If we are like you in the rest, we will [be identical to] you in that.
> (2) Therefore, I have authoritative rational intuitive justification for believing that, by living the life of a human minded animal, I am identical to a real human person.

In my opinion, Shylock's amazing words present a self-evident *Shakespearian* phenomenology of our "human, all too human" existence, in less than ten sentences. So, in view of the modal fact that classical identity is a necessary reflexive and symmetrical relation, and then reversing the direction of the conclusion in step (2), The Ecce Homo Argument provides authoritative rational intuitive justification for the belief that, as real human persons, we are *necessarily* human minded animal lives.

In opposition to both the classical Psychological Approach and also the classical Somatic Approach, including standard Animalism, and also by positively appealing to the evidential Ecce Homo Argument, what I want to argue is that the *only* view that captures all our authoritative philosophical rational intuitions about real personal identity is a *three-factor* approach, combining

> not only (i) biophenomenologically-based psychological elements,
> and also (ii) biologically/neurobiologically-based, somatic elements,
> but also a third factor, (iii) "freedom-dominance" (which I will describe in more detail immediately below),

as individually necessary, individually insufficient, and jointly sufficient conditions of real personal identity.

Now, let us think back to the first five chapters of this book, and to the Natural Libertarian metaphysics of free agency spelled out and defended there. Minded Animalism directly ties the Natural Libertarian metaphysics of free agency to the metaphysics of real personal identity, via the notion of freedom-dominance. Freedom-dominance says that only an animal life that is sufficiently deeply free (hence sufficiently filled with "ultimate sourcehood" and "up-to-me-ness") and deeply (non-)morally responsible, insofar as it is metaphysically embedded in the larger free-agency-structure that includes the capacities for veridical psychological freedom and principled authenticity, can be really *my own life*. So, in addition to the dual "Lockean" or two-factored biophenomenological-and-biological/neurobiological structure of personal identity, which combines both psychological and somatic elements, the third factor that imposes a further substantive necessary condition on the continuing life of real persons, is this—

> that *enough* of the dual biophenomenological and biological/neurobiological events in my conscious, intentional, caring, rational animal life must be also

> deeply free and deeply (non-)morally responsible, when embedded in the larger free-agency-structure that includes the capacities for veridical psychological freedom, and principled authenticity.

Otherwise, if *not enough* of the dual biophenomenological and biological/ neurobiological events in my life were deeply free and deeply (non-)morally responsible, then that life would belong ultimately either to an impersonal mechanistic nature, as the Conservation-laws-driven, Turing-computable movements of a natural automaton and/or to some external agency, as its mere causal effect and tool, namely, *to The Big Bang*, and therefore in neither of these cases *would my life really belong to me*. A naturally mechanized life and/or a Big-Bang-caused and Big-Bang-manipulated life, a life completely filled with non-veridical psychological freedom, would be nothing but *my so-called life* and not *my own life*.

To summarize so far. Minded Animalism is the combination of the four leading ideas listed above, into a single metaphysical doctrine of real personhood and real personal identity, namely,

> (1) the absolute, non-denumerably infinite, intrinsic, objective value of 2D rational animals or real persons,
> (2) the distinction between non-autonomous/lower-level/Frankfurtian real persons and autonomous/higher-level/Kantian real persons,
> (3) the tripartite criterion of real personal identity, composed of three individually necessary, individually insufficient, and jointly sufficient partial factors, namely,
> (3a) a biophenomenological factor, (3b) a biological/neurobiological factor, and
> (3c) a freedom-dominance factor, and
> (4) the Kantian/Existentialist thesis that a real person is identical to each and all of the proper parts or stages of her complete, finite, and unique 2D rational animal life-process.

This fourfold real-metaphysical menu makes it possible to showcase Minded Animalism's special theoretical virtues by critically comparing and contrasting it with Derek Parfit's influential reductionist theory of personhood and personal identity in *Reasons and Persons*, which I will do, step-by-step, in chapter 7.

6.2 REAL PERSONS

Here are some things I am accepting, with natural piety (see chapter 2 above), as primitive, phenomenologically self-evident, real-metaphysical starting points. You are a real person, and so am I. And so, I am assuming, is every other living organism that is capable of fully understanding, grasping, and feeling the normative force of these words. Neither logically possible or conceivable non-animal persons, disembodied persons, or divine persons, nor actual artificial persons (personae) or actual collective persons, created by human convention, are *real* persons in this sense. Real human persons are *essentially embodied minds*, hence they are persons who are non-logically, essentially non-conceptually, synthetically a priori, or "strongly metaphysically" necessarily, *human animals*. We all belong to a single universal intersubjective community of real human persons, and each one of us both inherently merits and also morally

requires equal consideration (if not the equal treatment[406]) and also moral respect for our human dignity, which entails that we morally owe everyone, including ourselves, especially those among us who are *oppressed*—that is, who are poor, or otherwise fundamentally neglected or mistreated, for example, by being coerced—whatever is *enough*, or sufficient, to satisfy the requirements of universal respect for human dignity, including, other things being equal, *never* treating people as mere means or mere things, and *always* providing people with what it sufficient to meet their basic human needs. We each have our very own complete, finite, and unique biological, neurobiological, first-personal, 2D rational lives. And at the same time we are all in this together, alive and living alongside one another in the selfsame fully natural and thoroughly nonideal, manifestly real world, for better or worse, till death do us part.

In a natural-pietist spirit, I believe that if those are not plain and simple facts, and also phenomenologically self-evident starting points for the real metaphysics of real persons, then nothing is. But those plain and simple facts, I also believe, also jointly confer on us a further special status, as I have already asserted several times: real persons are absolutely, nondenumerably infinitely, intrinsically, objectively valuable. Or to say the same thing with one word only: real persons have what Kant called *dignity* (*Würde*).

What, more precisely, do I mean by saying that? Objective values are whatever anyone can care about, that is, whatever anyone can aim her desire-based emotions at. Otherwise put, objective values are what Kant called "ends" (*Zwecke*). In turn, "absolute" means "unconditionally necessary." So to say that real persons, like us, are *absolutely, nondenumerably infinitely, intrinsically, objectively valuable*, or that they have *dignity*, is to say that *their value as ends is an unconditionally necessary, internal feature of the kind of being they are*.

Now many things are intrinsically objectively valuable, or ends-in-themselves—for example, pleasant bodily or sensory experiences, vivid emotional experiences, beautiful natural objects and environments, fine craftsmanship, skillfully-played sports, good science, good philosophy, good works of art, and any job well done. To say that real persons like us are absolutely, nondenumerably infinitely, intrinsically, objectively valuable, or that they have dignity, however, is to say that each of us has a moral value that is like a transfinite cardinal quantity in relation to all denumerable or countable, economic kinds of value. In accordance with this mathematical analogy, our moral value as real persons thus transcends every *denumerable* value quantity, and therefore every *economic* value quantity, yet remains fully in the natural world. As real persons, we are essentially *in* and *of* the natural world, but we are *not* "merely material" or "merely physical," in the sense of being strictly determined by the Conservation-laws-driven, Turing-computable, Big-Bang-caused, equilibrium or near-equilibrium thermodynamic and temporally reversible, naturally mechanized, non-purposive and non-teleological, non-living, non-minded, deterministic or indeterministic sub-parts of the natural world. Nor are we in any way *commodities*, which of course enables Minded Animalism to overlap significantly with early Marx's political theory, as formulated, for example, in the *Economic and Philosophical Manuscripts* of 1844. Any social institution or system that *commodifies* us, violates respect for our human dignity.

So yet again, the absolute, nondenumerably infinite, intrinsic, objective value of real persons is the highest possible kind of intrinsic objective value, sovereign over, or transcendent to, all other kinds; and the moral value of real persons cannot be provided either with an equivalent or anything greater in terms of any denumerable, economic value, commodity, or price—as it were, I'll trade you one Canadian-born independent philosopher, with a little over

150,000 miles on him, a few dents, and some rust, for a chilled can of Dale's Pale Ale, and the latest iteration of the iPhone, and throw in a new Lexus SUV as a bonus prize, and a named full professorship at Yale, Harvard, Princeton, NYU, Berkeley, UCLA, Stanford, Cambridge, or Oxford, with a $500,00.00 USD starting salary. That is all a joke, sort-of; but slavery is no joke, coercive authoritarianism is no joke, poverty is no joke, all forms of arbitrary prejudice leading to violations of respect for human dignity are no joke, and above all, treating either yourself or other people either as mere means to someone else's ends or as mere things is no joke. More generally, *human oppression* is no joke. Real human persons do not have a price, or a market value; real human persons are not commodities; the value of real persons is not merely instrumental; and more generally, real human persons cannot permissibly be merely used, abused, used up, or destroyed at will, and then thrown or flushed away.

To repeat, then, all real persons have *dignity* in this sense, and therefore all real *human* persons have *human dignity*. Above all, real human persons do not have to *do* anything in order to have human dignity, nor can they lose their human dignity *by acting badly*. Human dignity is neither *an achievement* nor *a reward for good conduct*: on the contrary, it is *an innate endowment*. Real human persons have absolute, nondenumerably infinite, intrinsic, objective value, or dignity, just in virtue of the innately specified and essentially embodied, more-or-less online[407] *capacities* they have for consciousness, intentionality, caring, 2D rationality, deep freedom, deep (non-)moral responsibility, and principled authenticity.

In this capacity-based way, real persons are what I call *subjects* of dignity, and *targets* of respect. *Other things being equal*, it is always morally impermissible to treat real persons merely as a means to some other end. Here I emphasize the "other things being equal" qualifier, aka the *ceteris paribus* clause, because, unfortunately, Kant did not adequately emphasize this with respect to his basic moral principles, insofar as they apply to 2D rational "human, all too human" animals in the thoroughly nonideal, manifestly real natural and social world.[408] Correspondingly, it is crucially important to recognize, that the dignity of real human persons does not entail that it is, strictly speaking and without qualification, *always* morally impermissible to treat people merely as a means to someone else's end. For example, under certain special contextual conditions, aptly captured by the famous (or notorious, depending on your methodological scruples in moral philosophy) thought-experiment of The Trolley Problem, it is clearly morally permissible to kill one innocent real human person in order to save five other innocents from being crushed by a runaway trolley, even by using that real human person *merely as a means*.[409] But on the other hand, *other things being equal*, manipulating real persons is immoral. Nevertheless, coercion—that is, manipulation either by means of violence or the threat of violence (primary coercion) or by means of non-violent salient harm or threats of non-violent salient harm (secondary coercion)—is indeed, strictly speaking and without qualification, *always* immoral. Correspondingly, the absolute, intrinsic, nondenumerable, objective value or dignity of real human persons also entails that it is, strictly speaking and without qualification, *always* morally impermissible to treat people *either as mere things, or in such a way as to rule out their actual or possible rational consent to that treatment*. In turn, these points about manipulation, coercion, and treating people as mere things, are not globally violated by our authoritative rational intuitions about The Trolley Problem.[410] Roughly, the idea is that permissibly killing one in order to save five in The Trolley Problem scenario, in that context, is the only really possible way of *protecting* and *saving* the five, and also of heeding a higher-order obligation *to choose the lesser evil in all such cases*, and *not* a way of coercing the one.

Otherwise put, and now fully in view of chapters 1 to 5 above, real persons are absolutely, nondenumerably infinite, intinsically, objectively valuable, essentially and precisely because they have the more-or-less online[411] complex, 2D rational capacity for *free agency*. So we have dignity essentially and precisely because we are *free agents*. As I have said many times already, the complex capacity for free agency (= free will + practical agency) is composed of the capacity for deep freedom and deep (non-)moral responsibility, which itself, as we have also seen many times already, is when a conscious, intentional, caring, rational animal's or real person's choices and acts

> (i) are really causally spontaneous,
> (ii) include the capacity for self-commitment to a live option, or Kierkegaardian Either/Or, and
> (iii) are owned by her, when and insofar as these factors are all also necessarily embedded in the larger free-agency-structure, including the capacities for veridical psychological freedom and principled authenticity.

It is also vitally important to note in this connection that it is *not* a general requirement of our having dignity—which, again, is essentially the same as our being capable of free agency—that we self-consciously recognize that *we ourselves* have dignity or are capable of free agency. Nor is it a general requirement of our acknowledging others as having dignity or being capable of free agency, that we self-consciously recognize that *they* have dignity or are capable of free agency. This is for two reasons.

First, the mental act or state of recognizing oneself or another real person as having dignity or being capable of free agency is not originally or primarily an act or state of self-conscious, or reflective, report, belief, or judgment. On the contrary, it is originally and primarily an act or state of pre-reflectively conscious emotional perception, or what Maiese and I call *affective framing*.[412] More precisely, on this view, emotional perception consists in an essentially embodied, conscious, caring intentional agent's representing the world *via her desire-based readiness to choose or act intentionally*, and, in the midst of that readiness, being disposed *to have feelings about the world, or others, or herself, in certain specific ways*; and the mental content of such acts or states of emotional perception is *essentially non-conceptual*.[413] These same points are also very effectively conveyed by Wittgenstein in the *Investigations*, without any technical terminology:

> "I believe that he is suffering." –Do I also *believe* that he isn't an automaton?
> It would go against the grain to use the word in both connexions.
> (Or is it like this: I believe that he is suffering, but am certain that he is not an automaton? Nonsense!)
> Suppose I say of a friend: "He isn't an automaton." –What information is conveyed by this, and to whom would it be information? To a *human being* who meets him in ordinary circumstances? What information *could* it give him? (At the very most that this man always behaves like a human being, and not occasionally like a machine.)
> "I believe that he is not an automaton," just like that, so far makes no sense.
> My attitude towards him is an attitude towards a soul. I am not of the *opinion* that he has a soul.[414]

Or in other words, we must accept the existence, presence, and dignity of real persons (in Wittgenstein's terminology, "human beings," or animals with "souls"), just as we must accept everything else that self-evidently appears in the manifestly real world, with natural piety.

Second, the necessarily equivalent concepts of DIGNITY and FREE AGENCY are characteristically *moral-metaphysical* philosophical concepts that are knowable or known only by rational reflection and authoritative rational intuition. It would be paradoxical in the extreme if, for example, someone's falling deeply in love and regarding another real person as inherently lovable required reflectively knowing the moral-metaphysical philosophical analysis of the concept LOVE, either partially or completely. On the contrary, obviously, romantic people normally affectively frame other people as inherently deeply lovable, and thereby fall deeply in love with them, without requiring any reflective or analytical grasp whatsoever of the concepts under which they themselves or the objects of their pre-reflectively conscious emotional perception fall. Correspondingly, it would be paradoxical in the extreme if, for example, someone's either being worthy of respect or respecting another real person required reflectively knowing the moral-metaphysical analysis of the concepts DIGNITY or FREE AGENCY, either partially or completely. On the contrary, people normally affectively frame themselves and others as having dignity, and as being worthy of respect for their own human dignity, and thereby respecting others and themselves, with pre-reflective natural piety, and without requiring any reflective or analytical grasp whatsoever of the concepts under which they themselves or the objects of their pre-reflectively conscious emotional perception fall.

The notion of our complex capacity for free agency—that is, deep freedom and deep (non-)moral responsibility, when and insofar as this capacity is necessarily embedded in the larger free-agency-structure that includes the capacities for veridical psychological freedom and principled authenticity—has three basic implications for the notion of personal identity.

First, if *P* is a real person, then *P* is deeply (non-)morally responsible for choosing or doing *X*—say, for drinking the last can of Diet Coke in the refrigerator, or for writing another 2000 words of her new novel—*if and only if P's choosing or doing X is deeply free.* Otherwise, choosing or doing *X* is not really and truly *her own* choice or act, which undermines her deep (non-)moral responsibility for choosing or doing *X*. In turn, *P* is deeply (non-)morally responsible for choosing or doing *X only if P is identical with the real person who chooses or does X*. Otherwise, it is not really and truly *P* who chooses or does *X*.

Now it is possible for *P* to choose or do *X without* being deeply free—for example, if *P* is unwittingly caused or manipulated to drink the last refrigerated can of Diet Coke, or write another 2000 words of her new novel, by a temporary mental illness, or by a powerfully manipulating evil scientist, etc.—and thus it is also possible for *P* to to choose or do *X*, without *P*'s being deeply (non-)morally responsible for choosing or doing *X*. But *P* is deeply (non-)morally responsible for choosing or doing *X if and only if P is identical with the real person whose choosing or doing X is deeply free.*

Second and more radically however, by way of the notion of freedom-dominance, I also want to claim that real personal identity is not possible *unless enough of the dual biophenomenological and biological/neurobiological states constituting the continuous life of that rational animal or real person are also deeply free and deeply (non-)morally responsible*. Framed in terms of psychological freedom, real personal identity is not *possible unless enough of the rational or real person's phenomenological states do exemplify veridical psychological freedom, and enough states do not exemplify non-veridical psychological freedom.* What precisely counts as "enough" will vary across real persons and contexts. For example, what

might count as freedom-overriding powerful manipulation, and therefore real personal identity-undermining powerful manipulation, in one rational human life—say, being raised by stern, tyrannical parents who are members of a religious cult that uses various kinds of behavioral modification techniques in order to to ensure social conformity, or having to endure a Stockholm-Syndrome-like social situation created by kidnappers or terrorists—might *not* in fact count as freedom-overriding and real personal identity-undermining powerful manipulation in a *different* real person's life. Some people simply have extremely strong, sharply self-defined, highly resilient personalities. But necessarily, for every rational animal or real person in any context, there will be some or another definite threshold that counts as a sufficient level of deep freedom and deep (non-)moral responsibility, and a sufficient level of veridical psychological freedom, for sustaining a distinctive real personal identity under those specific conditions.

Falling short of that context-sensitive threshold entails that this conscious, intentional, caring, animal is *not* actually living an acceptably 2D rational human life, that is, a human life that satisfies at least the Low-Bar standards of The Two Dimensional Conception of Rational Normativity—see section 3.1 above—but instead is merely *seeming* to do so. That is because this life is in fact relatively replete with non-veridical psychological freedom, and to that extent this life actually belongs either to an impersonal mechanistic nature, in the case of Natural Mechanism, or to some other external agency, in the case of powerful or freedom-overriding manipulation, and is not actually owned by the real person *herself*. It is something or someone *else's* feeling or doing, and therefore, catastrophically, the real person is merely a *tool* of that external agency. In the 1950s and 60s, they used to call this phenomenon *brainwashing*, although it now seems that this extreme version of what is nowadays more accurately and realistically called *mind-control*, was probably not real at that time. In any case, even if brainwashing *was* fictional and not real at the time, it remains *really possible* and in any case, effective mind-control techniques really did and do exist in cults, kidnapping and terrorism, military and police torture, and abusive personal and social relationships of all kinds.

A vivid fictional example from the movies is the Lawrence Harvey character, Raymond Shaw, in John Frankenheimer's brilliant 1962 black-comedy thriller, *The Manchurian Candidate*, who is an American soldier in the Korean War captured and brainwashed by the Red Chinese. Then he is "re-programmed" by a curiously jolly behavioral scientist to become a mindlessly obedient assassin—the behavioral equivalent of a *Terminator*—whenever he is presented with the triggering phrase, "Raymond, why don't you play a little solitaire?" And the *Robocop* movies are based on a similar premise, only with more high-tech machinery involved. Whether we call this overwhelming manipulation "brainwashing" or "effective mind-control," nevertheless, in all such cases, a real person's continued biological existence under those conditions is nothing but his "so-called life."

Third and perhaps most radically however, I also want to argue that if *P* is not deeply free when *P* chooses or does *X*, then *P* is not authentically identical with the real person who unfreely chooses or does *X*, *unless P is also prepared to take deep (non-)moral responsibility for some things over which she has no control*, especially including *the very desires or movements of her own body that were causally efficacious in P's choosing or doing of X, but were endogenously or exogenously caused, or powerfully manipulated by, something or someone other than P herself.*

What do I mean by this, and by the corresponding notion of *authentic identity*? I mean that if *P* has been endogenously or exogenously caused, or powerfully manipulated, to choose or

do *X*, then *P* is *not* deeply (non-)morally responsible for these very choices and body movements, and *P* subjectively experiences at most *non*-veridical psychological freedom with respect to choosing and doing *X*. That is because those so-called choices and doings were not deeply free, and correspondingly she did not enjoy veridical psychological freedom with respect to them: they merely *happened to her*. Hence those choices and body movements were not really and truly hers and did not truly belong to her own complete, finite, and unique life as a real person. They do remain fully within the dual biophenomenological and biological/neurobiological causal framework of her life as an individual 2D rational animal, but only as extrinsic to her free agency, and only as *opaque* to her intellectual and emotional self-understanding. In other words, those events remain as unresolved real personal and (non-)moral *problems* in her ongoing life. One vivid example would be Ketema Ross's violent actions under the grip of his schizophrenia; and another would be the subjective experiences of his elderly victims.

For convenience, I will call such events *Problematic Episodes*. Problematic Episodes generally defy a real person's intellectual or emotional self-penetration, and in specifically moral contexts, express a genuine dimension of *moral luck*. Moral luck, for my purposes here, is any state of affairs such that an agent can be treated as a subject of moral value, moral evaluation, and/or (deep) moral responsibility in relation to it, even though this state of affairs is significantly beyond the agent's control.[415] A Problematic Episode is an uncontrollable brute fact about an agent's life; yet, in specifically moral contexts, it also importantly exemplifies moral luck, in the following way. At any point later in time after the brute fact of the occurrence of unfreedom in a Problematic Episode, the real person can also deeply freely decide *to take deep moral responsibility for these unfree choices and body movements*, either by taking deep moral responsibility for the externally-caused or powerfully manipulated first-order desires that brought about those choices, or by taking deep moral responsibility for the externally-caused or powerfully manipulated biological/neurobiological states of affairs that brought about those body movements. If so, then she thereby freely appropriates those originally unfree choices and body movements, and in that sense *makes them morally her own*. She correctly remembers those subjective experiences *as* occurrences of non-veridical psychological freedom, and therefore as alienated from "who she really is"; but now she freely chooses to take them onboard, as proper parts of her own episodically-remembered past, even despite their having been significantly beyond her control. That is, she thereby brings them under the meta-control of her will, just by radically changing her attitude towards them, as a self-generated emotional *Gestalt*-switch. And in so doing, she personally resolves that Problematic Episode. This personal resolution produces essentially the same emotional and moral effect as Wittgenstein's Mystical Compatibilism (see section 4.1 above)—

> 6.43 If good or bad willing changes the world, it can change only the limits of the world, not the facts; not the things that can be expressed in language. In brief, the world must thereby become quite another. It must so to speak wax or wane as a whole. The world of the happy is quite another than the world of the unhappy.

Or in other words, the agent freely brings that Problematic Episode into "the world of the happy."

I will now generalize this important point to deep *non*-moral responsibility, in order to accommodate the by-now-familiar fundamental connection I see between free agency and

aesthetic/artistic activity—highlighted by Kant and Schiller in the late 18th century, and re-appropriated by Susan Wolf in the late 20th century—and frame it as a negative condition on personal identity. *Without* the rational animal's or real person's acts of free appropriation, whereby she takes deep (non-)moral responsibility for various natural facts, choices, or body movements that were not really her own to begin with—because they merely happened to her, and therefore because they were significantly beyond her control, or more specifically because they were caused or powerfully manipulated into existence by some agency other than herself—then there would be personally unresolved Problematic Episodes in her life. During such Episodes, she retains a dual biophenomenological and biological/neurobiological identity, and indeed also retains a real personal identity, providing that the freedom-dominance condition has been satisfied, *yet she lacks authentic identity*. So in this way, authentic identity is not only *a metaphysical relation*—by way of the three-part criterion of real personal identity—but also *a categorically normative fact*, by virtue of its being a partial achievement of principled authenticity. Or as a Sartrean Existentialist who is also a contemporary Kantian might put it:

> "At the end of the day, I am only whatever I manage to make of myself out of all sorts of antecedent brute facts and materials that I myself did not ask for and did not create, by wholeheartedly choosing and acting for the sake of the Categorical Imperative—and no excuses!"

That is, I am simply "condemned to be free" [416] in the three-leveled Natural Libertarianism sense of psychological freedom (and in particular, veridical psychological freedom), deep freedom, and principled authenticity.

If this perhaps surprising line of reasoning is correct, then since real person *P* is authentically identical with the real person who exemplifies deep freedom (ultimate sourcehood, up-to-me-ness) whenever *P* chooses or does *X*, it follows that a real person *P* is authentically identical with *all and only the deeply free choices and acts that P makes and performs*. And in this way, authentic personal identity can emerge as a *further* normative metaphysical fact from the *bedrock* metaphysical fact of biophenomenological, biological/neurobiological, and freedom-dominated real personal identity in all and only the creatures that are capable of psychological freedom (and in particular, veridical psychological freedom), deep freedom, and principled authenticity. This specifically includes all the *senior* rational human animals, namely, autonomous, higher-level, or Kantian real human persons. That is, it specifically includes all of *us*, the actual or possible critical, indifferent, or sympathetic readers of these words.

Am I saying that inauthentic people do not have real personal identities? No. But I *am* saying that insofar as we are inauthentic, *we are falling short of the kind of normatively rich personal identity we are all capable of achieving*, and that we should change our lives accordingly. Real personal identity is a categorically normative notion as well as a robustly metaphysical notion. We do not "construct" the real person or self, yet real personal or self-identity is *a life-project* carrying inherent norms of achievement/failure. We are morally or non-morally[417] required *not* to let ourselves slide into inauthenticity; and as the degree of inauthenticity increases towards a maximum, it gets closer and closer to being *non*-veridical psychological freedom, and therefore closer and closer to *non*-identity.

Now I need to face up more directly and explicitly to this question: What are "real" persons, as opposed to any other possible or actual kind of person—whether non-animal, disembodied, divine, artificial, or collective?

According to the Minded Animalist view I am now going to develop more explicitly, defend, and then critically deploy against Parfit in chapter 7 below, necessarily, all real persons are animals, but not all animals are real persons. Furthermore, necessarily, every real person is also an individual animal that inherently belongs to some species or another (for example, a real *human* person), but again the converse is not the case: not every individual animal within a species is a real person. For example, human infants born with anencephaly—without a cerebrum or a cerebellum, and lacking the top part of the skull—are really *biologically* human, but not real human *persons*. So not every individual human being is a real human person. And finally, not every particular living organism within a species is even *an individual animal within that species*, much less a real person in that species. For example, normal human embryos or zygotes prior to 14 days after conception, during the period of "totipotency," are not even individual human animals, precisely because during that period they can still either split into twins or fuse with several other embryos into a *chimera*. And as I mentioned above, as of 2018, new biomedical evidence suggests that many or even most women are totipotent chimeras during many decades of their lives.[418]

This account presupposes a certain kind of biological essentialism, and also a certain kind of realism about biological species. But on the account I favor, following Paul Griffiths's groundbreaking work on natural kinds, biological species-essences are *not* empirically hidden *intrinsic non-relational properties* of natural kinds—so biological species-essences are *not* noumena or "thing-in-themselves." Instead, they are manifestly real, intrinsic *relational* properties that necessarily include complex, self-organizing, far-from-equilibrium thermodynamics, and spatiotemporal asymmetry, and evolutionary historicity.[419] This "process structuralist" account of biological species-essences, in turn, is fully consistent with the immanent structuralist, neo-Aristotelian, and contemporary Kantian dynamicist philosophy of biology that I worked out in chapter 2 above.

It follows directly from my thesis that necessarily, all real persons are animals, that if God *were* to exist, precisely because God would be a purely spiritual agency, then necessarily S/He would *not* be a real person. Of course, I concede that it is logically, conceptually, analytically, or "weakly metaphysically" possible that God and other spiritual agencies, were they to exist, would be persons in *some* sense of the concept PERSON. But in view of the fact that I am doing *real metaphysics* in this book, for the purposes of my argument, I am justified in practicing *methodological eliminativism* with respect to all *spiritual persons*, by which I mean all disembodied or ghostly, infinite, immortal, omniscient, omnipotent, or omnibenevolent persons. Indeed, even those who insist on the personhood of God must also hold that S/he is a *radically different sort of person* from any real person, because this a direct consequence of the classical doctrine of the incomprehensibility of the divine mind. And how can a real person ever recognize an "Other Mind" that is an infinite, immortal, omniscient, omnipotent, or omnibenevolent person? The mystical doctrine of The Trinity is a failed attempt to respond to this powerful epistemic worry, but it seems clear that the only correct response is what I call *radical agnosticism*—and for much more on that, see *Kant, Agnosticism, and Anarchism*, part 1.

It also directly follows from my thesis to the effect that, necessarily, all real persons, including all real human persons, are animals, then necessarily, no machines can be real human

persons and no real human persons can be machines. So the very title of Julien Offray de La Mettrie's 18th century Cartesian materialist essay, *L'Homme Machine/Man a Machine*,[420] is, from the standpoint of real metaphysics, an oxymoron. Here the real-metaphysical situation is both interestingly similar to and also interestingly different from the case of God and other spiritual persons. It is interestingly similar to it, in that just as it is logically or conceptually possible that God or some other spiritual person, were it to exist, could also be a person in *some* sense of the concept PERSON, so too it is logically or conceptually possible that a machine could be a person in *some* sense of the concept PERSON. For example, science fiction is filled with intelligible and logically possible thought-experiments to this effect. But unlike the case of God and other spiritual persons, I cannot simply practice methodological eliminativism with respect to logically or conceptually possible machine-men or machine-women, for two reasons.

First, the term "machine" in English is ambiguous as between meaning

> either (i) a causally efficacious behavioral-functional-operational system that has been *artificially produced*, for example, in a factory or a laboratory,
> or (ii) a causally efficacious behavioral-functional-operational system that satisfies the necessary and sufficient conditions of Natural Mechanism.

It is naturally or nomologically possible that real organismic living systems could be artificially produced in a factory or a laboratory, as, for example, in Philip K. Dick's amazing classic sci-fi novel, *Do Androids Dream of Electric Sheep?*, and in Ridley Scott's correspondingly excellent classic sci-fi film, based on Dick's book, *Blade Runner*. But if so, then those systems would *not* satisfy the requirements of Natural Mechanism, since by hypothesis they are organismic living systems, not natural automata. So not all "machines" in sense (i) (= artificially-produced, causally efficacious, behavioral-functional-operational systems) are "machines" in sense (ii) (= natural automata).

Second, the sense in which it is possible for a person to be a "machine" in sense (i) is very different from the sense in which it is possible for a person to be a "machine" in sense (ii). It is *naturally or nomologically possible* for a "machine" in sense (i) to be a real person, by virtue of its being an artificially-produced living organism; and it is *logically or conceptually possible* for a machine in sense (ii) to be a person in *some* sense of the concept PERSON; but it is *non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" impossible* for a "machine" in sense (ii) to be a real person, since machinehood in sense (ii) constitutively rules out being-an-organism.

In any case, for the purposes of this book, and indeed for the purposes of THE RATIONAL HUMAN CONDITION as a whole, I am interested fundamentally and primarily in *real* persons, since, as I firmly believe, and as I am accepting with natural piety, that is what *we* are. Real persons are persons who are, "strongly metaphysically" a priori necessarily, minded animals. Real persons are essentially embodied, mortal, living organisms, and fairly limited in their knowledge, causal power, and goodness. Real persons are inherently open to good and bad luck alike. Real persons can, by virtue of their nature, be very happy, but can also suffer and be terribly unhappy. So real persons inherently exist in a *thoroughly nonideal* condition. Real persons cannot be "moral saints," if this means that they can be, even in principle, perfectly morally good. At best, real persons can achieve the sublime condition of real-world-saints or "sinner-saints," that is, real persons who more or less wholeheartedly strive to be as morally good as possible, yet still tragically fall short of that, given their thoroughly nonideal

condition.[421] Here, for example, I am talking about fictional or real-life people like Plato's Socrates; like Cervantes's Don Quixote; like Kierkegaard's Knight of Faith; like Joan of Arc as portrayed by Falconetti; like Dostoevsky's "Idiot" Prince Myshkin; like Watanabe as portrayed by Shimura; like Lincoln; like Ghandi; like Martin Luther King Jr.; and like Mother Teresa. To the extent that these fictional or real-life sinner-saints do inevitably fall short of their highest aims, and still make a mess of things to that extent, they suffer and cause suffering. Indeed, real persons, and especially real human persons, can even be quasi-defined as *the animals capable of evil (whether banal or near-Satanic) and suffering*.

What do I mean by the *suffering* of real persons? Here is an ostensive, phenomenologically vivid, display of what I mean, as expressed by Pablo Neruda and translated by Robert Bly[422]—

> Walking Around
>
> It so happens I am sick of being a man.
> And it happens that I walk into tailorshops and movie houses
> dried up, waterproof, like a swan made of felt
> steering my way in a water of wombs and ashes.
>
> The smell of barbershops makes me break into hoarse sobs.
> The only thing I want is to lie still like stones or wool.
> The only thing I want is to see no more stores, no gardens,
> no more goods, no spectacles, no elevators.
>
> It so happens that I am sick of my feet and my nails
> and my hair and my shadow.
> It so happens I am sick of being a man.
>
> Still it would be marvelous
> to terrify a law clerk with a cut lily,
> or kill a nun with a blow on the ear.
> It would be great
> to go through the streets with a green knife
> letting out yells until I died of the cold.
>
> I don't want to go on being a root in the dark,
> insecure, stretched out, shivering with sleep,
> going on down, into the moist guts of the earth,
> taking in and thinking, eating every day.
>
> I don't want so much misery.
> I don't want to go on as a root and a tomb,
> alone under the ground, a warehouse with corpses,
> half frozen, dying of grief.

> That's why Monday, when it sees me coming
> with my convict face, blazes up like gasoline,
> and it howls on its way like a wounded wheel,
> and leaves tracks full of warm blood leading toward the night.
>
> And it pushes me into certain corners, into some moist houses,
> into hospitals where the bones fly out the window,
> into shoeshops that smell like vinegar,
> and certain streets hideous as cracks in the skin.
>
> There are sulphur-colored birds, and hideous intestines
> hanging over the doors of houses that I hate,
> and there are false teeth forgotten in a coffeepot,
> there are mirrors
> that ought to have wept from shame and terror,
> there are umbrellas everywhere, and venoms, and umbilical cords.
>
> I stroll along serenely, with my eyes, my shoes,
> my rage, forgetting everything,
> I walk by, going through office buildings and orthopedic shops,
> and courtyards with washing hanging from the line:
> underwear, towels and shirts from which slow
> dirty tears are falling.

Of course, other kinds of minded animals can feel pain too, whether mild or intense. But there are significant differences in the quantity (that is, the total amount or degree) of pain on the one hand, and significant differences in the quality (that is, the specific character) of pain on the other, and they do not vary as mathematical functions of one another. Although every kind of minded animal can feel intense pain, not every kind of minded animal can feel the sort of self-conscious, self-reflective, richly content-laden, categorically normative, moral, emotional pain—namely, suffering—of which we are especially capable. Cats, dogs, horses, mice, cows, sheep, mice, and so-on, and so-forth, right through the non-rational minded animal bestiary, although they can feel varying degrees of pain, from almost unnoticeably mild to unbearably intense—*none* of them can *ever* feel the *kind* of pain so evocatively expressed by Neruda's poem.[423] Indeed, the so-called "superiority" of the human species pretty much boils down to the bracing natural fact that we are inherently more capable of *being desperately unhappy*, and also of *making each other desperately unhappy*, than any other animal species on Earth. *Ecce* the rational human animal.

If, necessarily, all real persons are animals within some species or another, then obviously we can make some headway on the question of the nature of real persons only if we are able to answer a preliminary question: "what is an animal?" The Oxford English Dictionary tells us that the word 'animal' means "a living organism which feeds on organic matter, usually one with specialized sense organs and nervous system, and able to respond rapidly to stimuli."[424] In the usage of contemporary biologists, the term 'animal' also has a taxonomical sense, in that animals are said to constitute one of the five kingdoms of living things: Monera (bacteria), Protists, Fungi, Plants, and Animals. The class of animals that is jointly specified by these ordinary language and biological-taxonomical senses includes vertebrates and invertebrates,

mammals and non-mammals—including birds, reptiles, amphibians, various kinds of fish, insects, and arachnids.

My usage of the term "animal" throughout this book, however, is a slight precisification of the ordinary language and biological-taxonomical usages, intended also to coincide with its use in *cognitive ethology*, that is, the scientific study of animal minds and especially non-human animal minds in the context of macrobiology, cognitive psychology, and behavioral psychology.[425] To signal this precisification, I have coined the quasi-technical term *minded animal*. Minded animals, as I have said, are conscious, intentional, caring living organisms.

Maiese and I argued in *Embodied Minds in Action* that necessarily every creature with a consciousness like ours is *an essentially embodied mind*.[426] Essentially embodied minds are the same basic kind of mind as ours, hence they are "minds like ours." An essentially embodied mind, or a mind like ours, in turn, is an irreducible consciousness that is also necessarily and completely neurobiologically embodied. This is to say that its irreducible consciousness cannot be disembodied, and that it thereby has a full-scale neurobiological incarnation of its conscious states in all its vital systems and vital organs—including the higher brain, brain stem, limbic system, nervous system, endocrine system, enteric system, immune system, and cardiovascular system, right out to the skin. So its consciousness is not only non-reducible but also non-dualistic and non-supervenient. Furthermore every consciousness, as the consciousness of an essentially embodied mind, or mind like ours, is fundamentally manifest as desire-based emotion, and, in particular, as effective desiring—which is the kind of desiring that is also either an effortless or effortful trying that causes intentional action.[427] So essentially embodied minds, or minds like ours, are always inherently poised for trying to *do* something, and thereby always inherently have a capacity for free agency.

The crucial idea of essential embodiment needs to be further elaborated. To say that every animal that has a consciousness like ours thereby has a full-scale neurobiological incarnation of its irreducibly conscious states in all its vital systems and vital organs—including the higher brain, brain stem, limbic system, nervous system, endocrine system, enteric system, immune system, and cardiovascular system—right out to the skin, is what Maiese and I called *The Essential Embodiment Thesis*.

It is important to note that the Essential Embodiment Thesis has two logically distinct parts:

> (1) the *necessary* embodiment of conscious minds like ours in a living organism (The Necessity Thesis), and
> (2) the *complete* neurobiological embodiment of conscious minds like ours in all the vital systems, vital organs, and vital processes of our living bodies (The Completeness Thesis).

The Necessity Thesis says that necessarily, conscious minds like ours are alive. Negatively formulated, it says that conscious minds like ours cannot be dead, disembodied, or naturally mechanized. By contrast, The Completeness Thesis says that conscious minds like ours are fully spread out into our living bodies, necessarily *including* the brain, but also necessarily *not restricted to* the brain. Please note that I am not saying that our brains, hearts, livers, or stomachs are themselves conscious. On the contrary, according to my view it is only whole animals and real persons that are conscious, not their body parts alone, and not even their brains alone. So what I am saying by asserting The Completeness Thesis is that the minded animal as

a whole—for example, a rational human animal, or real human person—is conscious *with*, or *in-and-through*, its brain, heart, liver, stomach, or whatever, right out to the skin.

One could, at least in principle, assert The Necessity Thesis and also reject the Completeness Thesis. And at least in principle one could also assert The Completeness Thesis and reject the Necessity Thesis—although it is somewhat harder to see what the point of asserting Completeness and rejecting Necessity might be, than the converse. But in any case I want to assert both Necessity and Completeness together. So I hold that the supposed consciousness of a causally detached brain—say, a living brain floating listlessly in a vat, as in Hilary Putnam's famous thought-experiment[428]—even though it seems both logically and conceptually possible, just would not be a consciousness *like ours*. On my view, a consciousness like ours non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" necessarily involves a brain that is causal-thermodynamically coupled with all the other vital systems, organs, and processes of our living body.

The notion of a "causal-thermodynamic coupling" is crucial. The Necessity Thesis and The Completeness Thesis do not jointly entail that a consciousness like ours actually is, or ever could be, embodied in *any* causally necessary condition of our kind of consciousness, which would of course include all sorts of entities and facts outside our living bodies. That is what Maiese and I call *The Embodiment Fallacy*.[429] Instead, the Necessity and Completeness theses jointly entail that consciousness like ours is embodied only in a special kind of fully-integrated far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic thermodynamic system that is both causally necessary and causally sufficient for consciousness like ours—namely, one that has all the same causal powers as the vital systems, organs, and processes of our living bodies. Any such living body is what we call *the natural matrix*, or natural basis, of a consciousness like ours.

And that point in turn raises another extremely important point, specifically about the very idea of a "natural matrix." A natural matrix of a consciousness like ours is not merely *a compositional material substrate*—a mass of physical stuff and a collection of physical parts—that necessarily accompanies and supports consciousness like ours. A natural matrix is, over and above that, a system of causal-thermodynamic relations, embedded in some or another compositional material substrate, awaiting specific activation or actualization. This means that if you significantly modify the shape of your body, or lose a limb or some other body part, without also replacing it with an equivalent counterpart that has the same relational causal powers, then you would also correspondingly modify or lose your mind. But the specific bodily stuff and the particular body parts are not metaphysically important.

What I am saying, then, is that the natural matrix of consciousness like ours is not just a hunk of specific bodily stuff, and not just a heap of particular bodily parts. Instead, the natural matrix of a consciousness like ours is all the vital systems, organs, and processes of our living bodies, as individuated by their relational causal powers—that is, as individuated by what they can efficaciously do in causal community with each other and with the larger surrounding world. That these vital systems, organs, and process are in fact composed of some or another hunk of specific bodily stuff and also of some or another heap of particular bodily parts—say, specifically human body stuff and particular human body parts—is of course extremely practically important for members of the relevant species made out of that stuff and those parts. But otherwise, it is metaphysically trivial. Thus The Essential Embodiment Thesis is a thesis about *the operative neurobiological non-equilibrium thermodynamics* of creatures with

consciousness like ours, and not, except trivially, a thesis about our compositional material substrate.

Assuming, then, that The Completeness Thesis is formulated in terms of the relational causal powers of the vital systems, organs, and processes of our living bodies, and not, except trivially, in terms of their compositional material substrate, there are at least four good reasons for defending The Essential Embodiment Thesis.

First, it seems obvious that if any of the vital systems, organs, or processes in our bodies is destroyed or permanently disabled without a functional replacement that has essentially the same relational causal powers—say, an artificial heart, a liver transplant, etc.—then our consciousness will cease to exist, precisely because the whole organism dies. Therefore *the existence* of consciousness like ours non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" necessarily depends on its complete neurobiological embodiment.[430]

Second, it seems equally obvious that significant changes made to the relational causal powers of any of our vital systems, organs, or processes normally produce correspondingly significant changes in the specific character of conscious minds like ours. And this is as true of the non-brain systems as it is of the brain systems. A thyroid gland malfunction, hormone imbalance, adrenaline surge, or heart attack is apt to cause highly significant changes in consciousness like ours. Therefore, *the specific character* of consciousness like ours also non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" necessarily depends on its complete neurobiological embodiment.[431]

To be sure, other things being equal, a lobotomy or a concussive blow to the head is apt to cause more fundamental changes in consciousness than a moderate thyroid malfunction, hormone imbalance, and so-on. And again, to be sure, the brain is centrally causally involved in every aspect of normal attentive, singly-focused, alert, self-conscious, self-reflective, waking consciousness. So I am not in any way denying the necessary and central causal role of the brain in the constitution of normal attentive, singly-focused, alert, self-reflective, waking human consciousness and intentionality. But at the same time, I am also strongly recommending that philosophers of mind and cognitive neuroscientists should not overemphasize the causal role of the brain, to the extent that this undermines our recognition of the equally necessary role of the relational causal powers of the rest of our vital systems, organs, and processes.[432]

For example, philosophers and cognitive neuroscientists could reflect, by way of comparison and contrast, on the fascinating case of the *octopus*, a minded animal whose mind is almost *literally* spread out all over its body—insofar as its body is almost entirely *arms*, and the majority of the neurons in its body exist outside its brain.[433]

And to take another example, closer to home, as everyone knows, even fairly minor changes in our *digestive* processes can produce non-trivial changes in consciousness. Think of the striking phenomenological differences between

> (i) feeling very hungry and craving a plate of spaghetti,
> (ii) feeling as if you ate just the right amount of spaghetti, and
> (iii) feeling utterly stuffed with spaghetti.

The brain obviously is centrally causally involved in these normal attentive, singly-focused, alert, waking phenomenological differences, but it seems also equally obvious that the brain does not in and of itself causally control or determine these differences. On the contrary, it

seems obvious that the "enteric brain"—our guts—is doing much of the causally controlling and determinative work here.[434] And similar points can be made about the other non-brain vital organs, systems, and processes. Each of them can and does play a causally controlling and determining role with respect to some differences in normal attentive, singly-focused, alert, self-reflective waking consciousness, even if the brain is also centrally causally involved.

Analogously, even if every basic act of a corporation passes directly through its Chief Executive Officer, it does not follow that the CEO controls or determines the specific character of every such act, or even most of them. In fact, in a great many cases the CEO is just the chief executive slave of the controlling determinations of the shareholders (if it is a public company), or of the employees (if it is either an employee-owned company or unionized), or of the actual business operations of the company. So too the brain is often just the central causal slave of the rest of the living body.

Third, there is empirical evidence in cognitive neuroscience that supports The Essential Embodiment Thesis. For example, recent work on the neurochemistry of human emotions strongly suggests that the vital systems centrally causally involved with and embodying our basic emotions are gut-based, not brain-based.[435] And recent work in cognitive psychology strongly suggests that human and non-human animal cognition cannot be adequately understood without taking into account the special neurobiological and environmental conditions of its necessary and complete non-brainbound embodiment.[436]

But fourth and finally, probably the most compelling empirical evidence for Essential Embodiment, precisely because it is the simplest, is the well-known fact that the "arc" of reflex action (say, someone's pulling her hand away from something very hot) operates more quickly than the time it takes for the brain to process information sent to it via the nervous system about the body parts involved in that reflex action (say, that the subject's hand has been seriously burned). If I am correct,[437] then this is also a conscious experience, although not of course a self-conscious or self-reflective experience: in the example of the burned hand, the subject's hand moves before she self-consciously or self-reflectively feels the searing pain of a burn. But I do think that reflex action still has a special phenomenology, in the classical Nagelian senses that there is a definite something-it-is-like-to-be for, and having-a-particular-point-of-view for, a suitably neurobiologically complex living organism like us when, for example, that organism is pulling her hand away from something very hot even though the self-conscious or self-reflective awareness of the searing pain of the burn has not yet emerged. If I am correct, then reflex action non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" necessarily includes a first-order and reflexive pre-reflective consciousness, even if it does not necessarily include a higher-order and self-conscious or self-reflective consciousness.

In addition to this point about pre-reflective or first-order consciousness, a further reason to think that reflex action is indeed reflexively conscious, although in a pre-reflective or first-order conscious way, is that it is possible to train oneself, through biofeedback strategies, to modulate or even suppress such reflexes.

So if this point about pre-reflective or first-order consciousness is correct, and if we also take biofeedback data seriously, then necessarily in cases of reflex action a pre-reflective or first-order consciousness occurs with and in-and-through the vital systems that constitute and subserve our intentional body movements, even though by hypothesis the brain is not centrally causally involved in the production of these pre-reflectively conscious intentional actions. Or in other words, there is compelling empirical evidence that there is a necessary and complete

neurobiological embodiment of consciousness even when the brain is only peripherally causally involved.

Now minded animals are always creatures within some real species *S* or another, hence they are always *S*-type (say, human, or feline, or canine, or equine, etc.) animals. But as I noted above, not every living organism within a species *S* is an individual *S*-type animal. For example, a single human embryo or zygote (that is, the sperm-fertilized ovum) is a living organism within the human species, in the strictly phylogenetic sense of sharing our species-specific biological essence, but a single human embryo is not necessarily a human individual. This is because, as I also noted above, early human embryos up to about the 14th day of their existence are totipotent. This means, among other things, that one embryo can split and later become two distinct human individuals (twins), and also that two embryos can fuse and later become a single human individual (chimeras).[438] Moreover, as I also mentioned above, it seems that many or most women are in fact totipotent chimeras.

What, more generally, is an *individual* belonging to some species *S*, that is, what is an *individual S*-type animal? My claim is this:

> Something *X* is an individual *S*-type (human, feline, etc.) animal if and only if *X* is a living *S*-type organism, and *X* is past the period of totipotency for that species *S*.

This defines an individual within a species in biological essentialist terms, although, as I noted above, this is only a *manifest* or causal-thermodynamic essentialism, a process-structuralist essentialism, and not a *noumenal* essentialism. More strikingly, perhaps, it follows from this conception of individuality-within-a-species that many or even most women are not individual human animals until fairly late in their lives, decades after achieving real personhood.

Within the human species—and also within a few non-human animal species—many or even most of the animals within that species can also become real persons within that species. The beginning of a real person's life for a given *S*-type animal is what I call the *neo-personhood* of that animal.[439] In the human species, as far as we currently know, the capacity for consciousness first manifests itself in normal fetuses between 25 and 32 weeks after conception or fertilization, hence roughly at the beginning of the third trimester.[440] My view is that this is when your very own life started—when you became a human neo-person. Prior to that, and from roughly 14 days after your parents conceived the human organism that eventually became you, there also existed a living human animal that also eventually became you—but, just like the totipotent human organism that became that human animal after 14 days, it was not yet you.

This distinction between *animals within a species S* on the one hand, and either *neo-persons* or *actualized real persons within a species S* on the other hand, is a deeply important difference, both metaphysically and morally. This can be seen in at two ways, with specific application to humans.

First, normal human fetuses after the period of totipotency but still before the emergence of consciousness at 25-32 weeks after conception or fertilization, are human animals *but not* real human persons, whether neo-persons or actualized real persons.

Second, anencephalic human infants[441] are human animals, but neither human neo-persons nor actualized real human persons.

Obviously these two claims, if true, will have serious implications for the morality of abortion and infanticide.[442]

Now here is a different point about minded animals. According to Minded Animalism, necessarily every *S*-type animal that is minded has a minimally coherent consciousness. But the minimal coherence of an animal's consciousness does not necessarily imply the unity of its consciousness. By the notion of a "unity of consciousness," I mean a conscious, intentional, caring animal's capacity simultaneously to *combine* or *comprehend* all currently experienced phenomenal characters or representational contents within a single phenomenological field. As I am using these terms, "combination" (roughly corresponding to what Kant calls "intellectual synthesis" at *CPR* B 151) is explicit, occurrent, structural awareness of characters or contents, whereas "comprehension" (roughly corresponding to what Kant calls "figural synthesis" also at *CPR* B 151) can involve implicit, dispositional, and substructural awareness. For example, attentively visually perceiving a red square next to a green square is an instance of a unity of consciousness in the mode of combination; but by contrast, the visual perception of a tree does not usually also include any explicit or occurrent awareness of any of its distinct leaves, even though these leaves are still perceived in a single *Gestalt* and might even be recalled—hence in such cases these visual contents are unitarily comprehended but not unitarily combined.

Over against the notion of a unity of consciousness—whether in the mode of combination (intellectual synthesis) or in the mode of comprehension (figurative synthesis)—by the notion of a "minimal coherence of consciousness," I mean any integration of a conscious, intentional, caring animal's phenomenal characters and representational contents such that the minded animal is capable of effective cognitive or practical intentional agency. Minimally coherent consciousness is perfectly consistent with a non-trivial amount of disunity within a minded animal's subjective experiences. For example, I can experience dissociated or divided cognition, but nevertheless retain a minimally coherent consciousness, even if for some reason—for example, neo-commissurotomy, that is, the recent surgical severing of the primary neurobiological pathway, or *corpus callosum*, that connects together the two halves of the higher brain—it is then neurologically impossible for me to bring together all my currently subjectively experienced characters or representational contents within a single conscious phenomenological field by means of any sort of combination or comprehension. And what *we* subjectively experience under very special conditions like neo-commissurotomy, the *octopus* must subjectively experience as its ordinary way of life.

In any case, according to Minded Animalism's criteria for being an *S*-type animal and for being an *S*-type real person, then even despite having "split brains," neo-commissurotomy patients still each possess a minimally coherent consciousness and therefore are unique real human persons. This claim is supported by behavioral evidence, which shows that most neo-commissurotomy patients exhibit "complete normalcy in ordinary activities," despite their cognitive dissociation, as Nagel pointed out in one of the earliest philosophical discussions of split brain phenomena.[443] Furthermore, in other cases in which the corpus callosum never actually develops—a phenomenon known as "agenesis" of the corpus callosum—compensatory sets of interhemispheric neurobiological pathways are formed in the brain stem, and adult subjects perform very close to normal on all psychological tests, including the tests for cognitive dissociation.

It is true that some neo-commissurotomy patients do develop pathological dissociations, for example, the famous case of the neo-commissurotomy patient whose hand would spontaneously smack his wife while he was simultaneously sincerely insisting that he loved her. But this case seems very close to "alien hand syndrome" and more generally to schizophrenic delusions of bodily control,[444] and if so, it would imply both a temporary non-

coherence of consciousness and also a disruption of deep freedom and moral responsibility with respect to these body-movements; and in this way it would constitute a set of Problematic Episodes in the wife-smacking man's life—and of course in his unfortunate wife's life as well.

By contrast to the wife-smacking man, a victim of Dissociative Identity Disorder, or DID (aka "multiple personality disorder")—assuming that the diagnostic evidence for this controversial condition is well-founded—would have a persistently incoherent consciousness and therefore would not be capable of effective cognitive or practical agency. Each distinct personality within a victim of DID is itself, perhaps, minimally coherent while it lasts; yet over time there is no single minimally coherent pattern of cognitive or practical agency. The separate pseudo-selves, it seems, are cognitively and practically alienated from one another. In this way the victims of DID, like other human minded animals who are permanently *non compos mentis* or incurably insane, and therefore persistently incapable of effective cognitive or practical agency, would be individual human minded animals but not actualized real human persons.

A very important implication of Minded Animalism, which follows from The Essential Embodiment Thesis, together with the explicit definition of real personhood that I am going to spell out in section 6.3 below, is that non-logically, essentially non-conceptually, synthetic a priori or "strongly metaphysically" necessarily, every real person has one and only one living animal body, and conversely, necessarily every living animal body of a real person (that is, every living animal body which is *also* lived by a real person—there can of course be living animal bodies which are not lived by real persons) is lived by one and only one real person.

Here is an argument for that claim. The Essential Embodiment Thesis entails that for each consciousness like ours, there is at least one living animal body; and every consciousness like ours is such that, when it is taken together with its corresponding living animal body, they *jointly hylomorphically constitute* exactly one minded animal. So for every minded animal that is also a real person then there is exactly one living animal body that is *lived* by that person. Conversely, if any living animal body is such that, when it is taken together with a corresponding consciousness like ours, they jointly hylomorphically constitute exactly one minded animal, and if that minded animal is also a real person, then there is exactly one real person who *lives* that living animal body. Therefore, non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" necessarily, every real person has one and only one biological/neurobiological life and necessarily every real person's biological/neurobiological life—that is, every biological/neurobiological life which is also the life of a real person—is continuous with one and only one real personal life. It is a bit of a mouthful, but for clarity's sake I will dub this *The One-Living-Body-Per-One-Real-Person-and-One-Real-Person-Per-One-Living-Body Thesis*.

This mouthful of a thesis entails, for example, that dicephalus (literally, "two-brained") twins—that is, so-called Siamese twins, human animals with two distinct brains and two distinct centers of consciousness, but who share the rest of their vital organs[445]—constitute two distinct living human animals that just happen to have a partial causal, material, and spatiotemporal overlap.

Quite generally speaking then, according to Minded Animalism, no matter what species *S* you belong to, you are numerically the very same *S*-type real person *from the very moment* you acquire the complete biological/ neurobiological embodiment of your consciousness, thus entering the stage of your neo-personhood, or the beginning of your life, providing that enough of the events in that later life are deeply free and deeply (non-)morally responsible, insofar as this is metaphysically embedded in the larger free-agency-structure that includes the capacities

for veridical pyschological freedom and principled authenticity. Then you continue to be numerically the very same *S*-type real person throughout your life as long as the complete neurobiological embodiment of your single, minimally coherent consciousness continues to exist, and again providing that enough of the events in that life are deeply free and deeply (non-)morally responsible within the larger free-agency structure that includes the capacities for veridical psychological freedom and principled authenticity, until the very moment the complete neurobiological embodiment of your consciousness is destroyed or permanently disrupted, and you die.

But insofar as you ever *fail* to satisfy those conditions, and even whether or not you are at that particular time capable of principled authenticity, at least partially or to some degree—for example, if you happen to be a normal human toddler or older child, or a non-human animal real person—then you are *not* numerically the very same *S*-type real person.[446] Or in other words, an *S*-type real person is not merely *alive*: it also *has a life* or *is the subject of a life*, which is to say that it is also the egocentric center of a single, minimally coherent, essentially embodied, living organismic, conscious, intentional, caring, 2D rational, complete, finite, and unique freedom-dominated *S*-type life.

I will formulate the Minded Animalism criterion of personal identity more carefully and formally in chapter 7. But for the time being, it can be sufficiently clearly seen that the Minded Animalism criterion directly implies that the unfortunate Terri Schiavo was *not* the very same human animal both before and after her catastrophic heart attack, precisely because after her catastrophic heart attack the individual real human person that she had been, in fact no longer existed. So Terri Schiavo's single, coherent, essentially embodied, living organismic, conscious, intentional, caring, 2D rational, complete, finite, and unique freedom-dominated human life ended with her heart attack—that is, catastrophically, her life as a real human person then ended—although a particular living human organism that was causally and biologically/neurobiologically continuous with her body survived and existed in a persistent vegetative state for 15 years afterwards. Thus a human organism called "Terry Schiavo" remained alive for 15 years, but Terry Shiavo herself no longer had a life and no longer was the subject of a life.

6.3 Necessary and Sufficient Conditions for Real Personhood

This brings me to the explicit real-metaphysical analysis of the concept and fact of a real person. Again, according to Minded Animalism, every real person is also an *S*-type animal or living organism (but not conversely), and every *individual S*-type animal is also an *S*-type animal or living organism (but not conversely). Therefore, being an *S*-type animal or living organism (although not necessarily an *individual* one, in order to accommodate totipotent organisms in general and chimeras in particular) is a necessary although not a sufficient condition of real personhood. The rest of my real-metaphysical analysis substantively borrows from two different sources:

(i) Harry Frankfurt's hierarchical-desire theory of persons, and
(ii) Kant's rationality-based theory of persons.

As I just indicated, and perhaps at first unintuitively, Frankfurt's theory of persons is based on the notion of an hierarchically-structured set of desires. The fundamental connection here is that for Frankfurt, a person is essentially *identified with* the constitution of her will, which in turn is a set of desires immanently structured by the capacities for rationality and free agency, and inherently governed by the norm of "decisive identification wth effective first-order desires," that is, by the norm of authenticity or wholeheartedness. In a nutshell, *that is my view of real persons too*, although with an explicitly 2D conception of rational normativity; and with a more explicitly and robustly Kantian twist, or rather, set of twists; and also with my mereological conception of a personal identity-relation operating over whole lives and proper parts of those lives. Freedom and moral responsibility on my account are *deep* freedom and *deep* (non-)moral responsibility, and deep freedom is locally incompatibilistic, whereas Frankfurt's account is explicitly friendly to classical Compatibilism/Soft Determinism. Moreover, the guiding norm of real personhood, according to my account, is the High-Bar categorical norm of principled authenticity, whereas Frankfurt's account is not explicitly committed either to categorical (as opposed to merely instrumental) norms of agency and personhood or to overriding-reasons-providing moral principles in the Kantian sense.

Even so, bracketting the differences I have just highlighted, I do think that the full sweep of Frankfurt's work, especially including the essays included in *The Importance of What We Care About*, published in 1988, and also his later 2004 book, *The Reasons of Love*, is largely consistent with and theoretically friendly to my commitments to categorical norms of agency and personhood, and also to my commitments to overriding-reasons-providing moral principles in the Kantian sense: so that is the working assumption I will make as I press forward.

But let me now explore some further specific Frankfurtian details, because they are fundamentally important for my account of real persons. As I mentioned in section 3.2, on my view a desire is a felt need for something, or a conscious going-for something. This is as opposed to an *actual* need for something—obviously not all felt needs are actual needs—and also as opposed to a mere pro-attitude towards something, a mere preference for something, or a mere wish for something. Frankfurt himself defines the notion of a desire somewhat more broadly, so as to include all pro-attitudes, preferences, and wishes; but in the present context, it is convenient to use my narrower and more conative notion of a desire. Desires in this sense are essentially equivalent with active, committed wants. So to desire X is actively and committedly to want X; and to desire to X is actively and committedly to want to X.

According to Frankfurt, some animals have not only what he calls *first-order desires*, which are ordinary direct desires for things, events, or real persons (for example, the infant wanting her mother), but also *effective first-order desires*. Effective first-order desires are desires that move (or will move, or would move) the minded animal all the way to action. An effective first-order desire is the same as a minded animal's *will* or *first-order volition*. First-order desires may or may not be accompanied by *second-order desires*: to want (not) to want X, or to want (not) to want to X. If so, then some of the second-order desires may be directed to the determination of precisely which first-order desire is to be the effective first-order desire, that is, the minded animal's will and first-order volition; and such desires are *second-order volitions*.

According to Frankfurt, whatever the order-level of desires or volitions, they can be either conscious or non-conscious. For the purposes of my discussion, however, I will concentrate exclusively on *conscious* desires and volitions. This is, in part, because I think that there is no

such thing as a mental state, whether dispositional or occurrent, that is strictly non-conscious and not to some non-trivial degree occurrently conscious. In earlier work, Maiese and I have called this (admittedly controversial, but also, we believe, defensible) claim "The Deep Consciousness Thesis."[447] But in any case, and according to Frankfurt, all and only persons have second-order volitions, because all and only persons care about the precise constitution of their wills. By contrast to persons, creatures that are "wantons" have effective first-order desires, but they either lack second-order desires (hence they cannot care about the precise constitution of their wills because they lack self-conscious desires) or if they have second-order desires they nevertheless lack second-order volitions (hence even though they have self-conscious desires, they still cannot care about the precise constitution of their wills). Again, according to Frankfurt, all non-human animals, all human infants, and some human adults are wantons. Finally, for Frankfurt a person has freedom of the will if and only if she can determine, by means of a second-order volition, precisely which among her first-order desires is the effective one. This is also known as *identification* or *decisive identification*;[448] otherwise persons have unfreedom of the will. Wantons have neither freedom of the will nor unfreedom of the will, simply because they are not persons.

As I said above, I accept much of what Frankfurt has to say about persons and their wills, and correspondingly I want to apply much of what he says to *real* persons and *their* wills. Nevertheless, I also have substantive disagreements with him on two mid-sized (as opposed to either major or minor) points.

My first mid-sized point of substantive disagreement is that I doubt that Frankfurt's notion of personhood adequately captures the full breadth or depth of my contemporary Kantian notion of real personhood, according to which some real persons have what I will call *higher-level or Kantian (2D) rationality*. This, in turn, is an innate complex capacity for strict-norm-guided logical or practical reasoning, for reflective self-consciousness, for autonomy or self-legislation, for authenticity or wholeheartedness, and for deep (non-)moral responsibility. A minded animal that also has higher-level or Kantian (2D) rationality can recognize necessary truths, judge or believe with a priori certainty, and choose or act wholeheartedly in accordance with desire-overriding non-instrumental, non-selfish, non-egoistic or non-self-interested, non-hedonistic, non-consequentialist, categorically normative reasons and duties, that is, those reasons and duties that inherently express the Categorical Imperative and the "categorical 'ought'."

By sharp contrast, what I will call *lower-level or Humean (2D) rationality* involves only the possession of innate capacities for conscious, intentional desire-based logical or practical reasoning, for more or less momentary or occasional occurrent self-consciousness, and for self-interested, or in any case instrumental, intentional agency. A minded animal that has lower-level (2D) rationality can recognize contingent truths, judge or believe with a posteriori certainty, and choose or act in accordance with broadly instrumental egoistic, hedonistic, or consequentialist reasons and duties, or those that express at most the "hypothetical 'ought'."

All minded animals that possess an innate capacity for higher-level or Kantian (2D) rationality also possess an innate capacity for lower-level or Humean (2D) rationality, but not the converse. For example, it is arguable that normal, healthy Great apes and perhaps also dolphins[449] possess an innate capacity for Humean or lower-level (2D) rationality, but not a capacity for higher-level or Kantian (2D) rationality. This is of course *not* to say that Great apes or dolphins are "irrational" or "non-rational" in any sense. On the contrary, it is only to say that, relative to those animals that *do* possess an innate capacity for higher-level or (2D) Kantian

rationality, the (2D) rational capacity of Great apes and perhaps also dolphins is somewhat limited in complexity and normative power. Minded animals with an capacity for (2D) rationality in the higher-level or Kantian sense are not only constrained in their intentional agency by the Categorical Imperative or at least by some strictly universal, non-instrumental, altruistic, non-hedonistic, and non-consequentialist moral reasons and objective principles, they are also capable of being moved wholeheartedly by the higher-order moral emotion of respect.[450] Or in other words, minded animals with a fully online capacity for (2D) rationality in the higher-level or Kantian sense are also capable of Kantian autonomy and principled authenticity.[451]

By contrast, minded animals that possess only an innate capacity for (2D) rationality in the lower-level or Humean sense are constrained in their intentional agency only by (at least some of) the axioms of rational choice theory, but not by strictly universal, non-instrumental, altruistic, non-hedonistic, and non-consequentialist moral reasons and objective principles. They are therefore not capable of Kantian autonomy or principled authenticity. Instead, they are at most capable of being moved non-authentically or non-wholeheartedly by the first-order Humean moral emotion of *sympathy*.[452]

What is the moral-emotional difference between Kantian respect and Humean sympathy? One way of cashing out this difference is to say that

> whereas (i) someone who is being moved by Kantian respect will always and necessarily choose and act so as to heed or preserve the dignity of another real person, even if she does not find that other real person to be *in any way whatsoever* attractive, likeable, nice, tear-jerkingly pathetic, or pleasant—in short, even if she involuntarily finds that real person to be perfectly loathsome,
>
> nevertheless (ii) someone who is being moved merely by Humean sympathy will choose and act so as to heed or preserve the dignity of another real person *only if* she finds that real person to be appropriately attractive, likeable, nice, tear-jerkingly pathetic, or pleasant.

In other words, mere Humean sympathy cannot survive the apparent loathsomeness of other real persons: *mere Humean sympathy* loses heart in the face of *involuntary disgust*. But Kantian respect inherently can and always does recognize dignity, even in the face of involuntary disgust.

—And this is *not* a superhuman, or "moral saint-like," moral attitude. For example, I imagine that a great many medical doctors all over the world, especially general practitioners, and nurses or nurse-practitioners, all of whom are "human, all too human," just like the rest of us, perfectly illustrate the sharp moral difference between respect and sympathy almost every single day of their working lives. Indeed, some, like Camus's fictional Dr Rieux in *The Plague*, and the real-life Florence Nightingale, illustrate this even to the sublime level of being real-world moral saints or "sinner-saints."

Some human animals are "persons" in Frankfurt's sense, hence are real persons in my sense, and also (2D) rational agents in the lower-level or Humean sense, and not (2D) rational agents in the higher-level or Kantian sense, but this is *not* because they lack the innate capacities for agency in the higher-level sense. Rather they *do* possess these capacities, but in the mode of *real potentiality that is not yet actualized*, hence it it is simply because they are *not-yet* (2D)

rational agents in the higher-level or Kantian sense. Indeed we—the actual and really possible readers of this book—were all of us, for a time, such creatures.

Consider, for example, normal *toddlers*. Normal toddlers are healthy human children between the ages of roughly 1 to 3 years who are just beginning to walk and talk. It is difficult to remember being a toddler, although of course almost everyone has a few memories from that period. But even if the introspective, memory-based phenomenological evidence is fairly thin, I think that any adult who has ever lived with and looked after a toddler knows that it is possible for someone to care very deeply about the constitution of her will, but not yet be capable of norm-guided logical or practical reasoning, self-reflection, or principled authenticity. Normal human toddlers do indeed manifest *moments* of non-instrumental, altruistic, non-hedonistic, and non-consequentialist thinking, feeling, and action. These are truly lovely anticipatory flashes of the higher-level or Kantian real persons they will eventually become, if their luck holds up, when all their basic innately-specified moral capacities are fully online—that is, they have fully matured, thereby becoming operative and well-developed.[453]

In any case, toddlers are conscious, intentional, caring animals, and they also are self-conscious in the minimal sense that they can recognize themselves in a mirror and make simple judgments about some of their own mental states. They are naturally affectionate and often highly sensitive to the changing emotional states of those around them. Usually, they have a minimal and rapidly-developing competence in their native natural language(s). They have simple beliefs, and if they have also acquired minimal linguistic competence, then they can carry out some simple inferences. They want things. Also they usually know what they want; they care a great deal about getting what they want; and, often enough, they can also determine which of their first-order desires is to be the effective one. So they have second-order volitions. They can mentally cause movements of their own bodies by means of their second-order volitions and effective first-order desires. They know their own names. They are intellectually curious, have capacious memories and wonderful imaginations, and are sometimes highly insightful. And for a few truly lovely moments at least, they can spontaneously think, feel, and act in non-instrumental, altruistic, non-hedonistic, and non-consequentialist ways.

But they are also extremely naïve and uncritical, highly emotional, highly inconsistent in their behavior, and above all characteristically self-centered, fickle, and willful, and certainly cannot be, or be held to be, deeply morally responsible for their actions. They need to be most carefully looked after, gently told what to do, loved unconditionally, and at all times protected from the vicissitudes of an often unfriendly and violent world. They are incapable of conceiving their own lives as a whole, and are therefore not self-reflective agents. They cannot engage in retrospective or prospective self-interpretation and self-evaluation, explicit step-by-step deliberation about immediate action, or long-term planning. They are certainly not capable of either deep moral or non-moral responsibility, or principled authenticity. And they certainly cannot grasp the moral significance of their own deaths.[454]

For these reasons, a toddler can, in a minimal way, self-consciously "identify" in the Frankfurtian sense—that is, make a decision—by using a second-order volition to determine precisely which among her first-order desires is her effective first-order desire. But the self-conscious "identifications" or decisions of toddlers are at best momentary or temporary, and do not occur consistently or over an extended period of time.[455] As I mentioned just above, toddlers cannot achieve or even comprehend either deep (non-)moral responsibility or principled authenticity. Toddlers are only momentarily moved to intentional choice and action by respect for human dignity, even if they often feel it inchoately, in an essentially non-

conceptual way. And toddlers are never synoptic thinkers, moral exemplars, or sublime real-world moral saints, "sinner-saints."

In this connection, toddlers and other young children as a class are very usefully compared and contrasted with adolescents or teenagers. Adolescents are somewhat capable of achieving principled authenticity at least partially or to some degree, but again still not fully so. In my terminology, they are most accurately characterized as *semi-Kantian* real persons. This is because adolescents are still significantly self-centered, fickle, and willful. They still cannot adequately grasp the moral significance of their own deaths. They still are somewhat incapable of authentic or wholehearted commitment to their own moral principles, the Categorical Imperative, and the dignity of real persons, including taking complete responsibility for some things over which they have no control. So they still cannot be fully moved by respect for human dignity.

But unlike toddlers and other young children, adolescents can be intensely passionate, utterly reckless, and deeply romantic. Shakespeare's *Romeo and Juliet*, Somerset Maugham's *The Razor's Edge*, Nicholas Ray's *Rebel without a Cause*, and J.D. Salinger's *Catcher in the Rye* all beautifully capture—in their very different ways—"what-it-is-like-to-be-an-adolescent," hence what it is like to be a semi-Kantian real person. As Salinger's Holden Caulfield so vividly saw, it is like living in a field of rye on the edge of a precipice—as it were, on the edge of Kierkegaard's chasm of 70, 000 fathoms—but at the same time also *intensely wanting to take deep moral responsibility* for catching young children before they stray too close to the edge, and fall helplessly into the abyss.[456] Switching over now from Kierkegaard's chasm to Maugham's lovely metaphor, adolescent persons are poised *on the razor's edge*, neither fully child nor fully adult. So there are substantive and not merely conventional reasons why we do not let adolescents marry, why we do not send adolescents to war, and why we do not treat them as fully morally responsible adults in courts of law.

Of course, as we all know, it is mostly not any of their own doing: it is, as the old saw has it, *mostly their hormones talking*. In any case, whatever the underlying physiological causes are, their basic innate rational-moral capacities are still somewhat immature, latent, and undeveloped. But the self-evident fact remains that adolescents, as a proper sub-class of normal human beings who are both psychophysically and morally distinct from toddlers and other young children on the one hand, and also psychophysically and morally distinct from fully autonomous and fully responsible adults on the other, are excessively self-focused, insufficiently capable of self-control, insufficiently self-reflective, insufficiently capable of long-term planning, and insufficiently capable of fully recognizing the dignity of other persons or of themselves. Somewhat like toddlers and other children, adolescents are almost never synoptic thinkers, moral exemplars, or sublime real-world moral saints, "sinner saints." For example, as many readers of Dickens have correctly pointed out, the lovely, gentle, preternaturally wise adolescent character Little Nell in *The Old Curiousity Shop* is mainly a reflection of Dickens's own kitschy, sickly-sweet fantasies about young women—although perhaps fewer have agreed with Oscar Wilde's wickedly witty remark that it is impossible not to laugh when poor Little Nell expires.

Be that as it may, adolescents are indeed capable of relatively sustained periods and phases of non-instrumental, altruistic, non-hedonistic, and non-consequentialist thought, feeling, and action, in anticipation of their full Kantian real personhood—again, as in the case of toddlers, a truly lovely thing. But sharply unlike toddlers, although very like most other older children,

normal adolescents are quite capable of doing some philosophy, and some of them are extremely good at the formal parts of philosophy, especially logic.

Of course, I am not neglecting the fact that mature normal rational human animals, or real human persons, who are fully capable of higher-level or Kantian rationality, and also principled authenticity, can sometimes be very self-centered, fickle, and willful too, just like toddlers or adolescents. But they are not characteristically or typically so. A normal human adult who characteristically or typically comported himself in the way that toddlers, other young children, or even adolescents characteristically or typically do, would, no doubt, be correctly regarded as an unfortunate victim of some cognitive or emotional syndrome, and treated accordingly.

My main point here is simply that normal toddlers and other young children are capable of a great many different sorts of complex affects and complex emotional states, including momentary anticipations of Kantian real personhood, but they are not (yet) higher-level (2D) rational human animals. Toddlers and other young children are therefore real human persons in the Frankfurtian sense, but not (yet) real human persons in the Kantian sense. Toddlers and other young children are therefore junior real human persons—as it were, tenure track assistant professors in The Realm of Ends—but not yet senior real human persons, as it were, tenured full professors in The Realm of Ends, even though they are, in terms of dignity, equal moral persons with all the other real persons. By another contrast, adolescents or semi-Kantian persons more generally are poised, as I have said, on a razor's edge somewhere between the higher and the lower levels. They are equal moral persons of the middle rank, but not card-carrying members of the classes above or below themselves—tenured associate professors in The Realm of Ends.

Thus real personhood in the lower-level, Frankfurtian sense is a necessary and sufficient condition of real personhood, which includes all the more-or-less online basic capacities of free agents, hence it entails dignity. And as we have seen, it is based on the fully online capacity for having second-order volitions, which in turn contains several other distinct constituent fully online psychological capacities. Real personhood in the higher-level, Kantian sense, on the other hand, both includes and significantly augments real personhood in the Frankfurtian sense, by including the fully online capacity for principled authenticity, at least partially or to some degree. Correspondingly, real personhood in the higher-level, Kantian sense is based on the fully online capacity for higher-level rational agency, which also contains several other distinct online psychological capacities.

In order to display the internal complexity of the relationships between these capacities more fully, here is an explicit version of the two-level theory of real personhood that I have been developing, in the form of the following tripartite definition.

> A Three-Part Definition of Real Personhood
>
> Part I. X is a real Frankfurtian person ($person_f$) if and only if X is an S-type animal and X has fully online psychological capacities for
>
> (i) essentially embodied consciousness or essentially embodied subjective experience,
> (ii) intentionality or directedness to objects, locations, events (including actions), other minded animals, or oneself, including cognition (that is, sense perception,

> memory, imagination, and conceptualization), and caring (that is, affect, desire, and emotion), especially including effective first-order desires,
> (iii) lower-level of Humean (2D) rationality, that is, logical reasoning (including judgment and belief) and instrumental decision-making,
> (iv) self-directed or other-directed evaluative emotions (for example, love, hate, fear, shame, guilt, pride, etc),
> (v) minimal linguistic understanding, that is, either inner or overt expression and communication in any simple or complex sign system or natural language, including ASL, etc., and
> (vi) second-order volitions.
>
> Part II. X is a real Kantian person ($person_k$) if and only if X is a real $person_f$ and also has fully online psychological capacities for
>
> (vii) higher-level or Kantian (2D) rationality, that is, categorically normative logical 2D rationality[457] and practical 2D rationality, the latter of which also entails a fully online capacity for deep (non-)moral responsibility, autonomy (self-legislation), and wholeheartedness, hence a fully online capacity for principled authenticity, at least partially or to some degree.
>
> Part III. X is a real person if and only if X is either a real $person_f$ or a real $person_k$; and any other finite, material creature or entity X is a non-person.

It is crucial to note that each one of the necessary conditions of real personhood is intended to be a "constitutively necessary condition" of real personhood, in the following sense, where *X* and *Y* are properties corresponding respectively to the concepts of *X*-hood and *Y*-hood—

> *X* is a *constitutively necessary condition* of *Y* if and only if
>
> (i) *X* is a necessary condition of *Y*,
> (ii) the existence and specific character of *Y*-facts is partially determined by the existence and specific character of *X*-facts, and
> (iii) the essence of *Y*-facts presupposes the essence of *X*-facts.

In this way, roughly, *X-hood flows from the essence of Y-hood*. Or in other words, the analysis of real personhood I am proposing is not a conceptual analysis of the concept REAL PERSON, but in fact a "real definition" and a contemporary Kantian real metaphysical analysis of the things that fall under that concept, given in terms of the essential properties of those things. Therefore, given that it is a true statement to say that (for example) a real person possesses a fully online capacity for essentially embodied consciousness, then this statement is a non-logical, essentially non-conceptual, synthetic a priori, or "strong metaphysical" necessary truth, not a logical, conceptual, analytic, or "weak metaphysical" necessary truth.

I will not stop here to discuss the analytic-synthetic distinction, although I do so in in depth elsewhere.[458] It would be a long stop. I will only note in passing that I accept the strongest version of the distinction, which I call *Kant's Pitchfork* (as opposed to "Hume's Fork"), because it includes not only the much-controverted Kantian "modal dualist" thesis that there are two irreducibly and categorically different types of necessary truth and non-empirical knowledge

(that is, the synthetic a priori and the analytic), but also the classical Humean and Logical Empiricist thesis that there is an irreducible and categorical difference between necessary truth and non-empirical knowledge on the one hand, and contingent truths or falsehoods and empirical knowledge (that is, the synthetic a posteriori) on the other.

Before going on to say more about real persons, there is a very important point I need to make about *non*-personhood. Just because a creature is a non-person, it does *no*t follow that this creature is thereby without *any* moral value, that is, a "mere thing." It is true that non-persons are neither subjects of dignity nor targets of respect. But at the same time, all living things and all minded animals—even what I call "minimally minded animals," for example, octopuses and other marine animals, insects, reptiles, and other invertebrates[459]—and all conscious or fully minded non-human animals, like bats, bears, birds, cats, dogs, horses, lions, mice, and wolves, are either *experiencers* or *subjects* of moral value, to some degree, and, correspondingly, to some degree, *targets* of our moral concern, even if they are non-persons. This is because they share with us three constitutively necessary conditions of real personhood, namely,

> (i) organismic life,
> (ii) to some degree, the capacity for consciousness or sentience, and
> (iii) to some degree, the capacity for free volition or animal agency,

all of which have intrinsic moral value, to some degree. But in any case, I have much more to say about these points in *Kantian Ethics and Human Existence*.[460]

Autonomy in the Kantian sense, as I construe it, and as I have already described it in chapters 3 and 5 above, is veridically psychologically free, deeply free and deeply (non-)morally responsible, self-conscious, self-reflective *self-legislation* according to the Categorical Imperative or moral law, the recognition of which provides an overriding non-instrumental reason for action, and also causally triggers the innate emotional disposition to feel respect for persons and the moral law, which in turn generates the wholehearted second-order desire to achieve moral self-transcendence. So Kantian autonomy in this sense, together with the capacities for wholehearted second-order desire for moral self-transcendence and the moral emotion of self-fulfillment, jointly constitute the capacity for achieving principled authenticity at least partially or to some degree. In its occurrent or realized version, principled authentity, in turn, is a person's wholehearted adherence to her moral principles and to some absolute moral principles, together with her sometimes taking deep (non-)moral responsibility, with no excuses, for things over which she had no control. In view of this latter feature, as I said in section 3.4, another illuminating label for principled authenticity would be *passionate Kantian stoicism*.

A "self-reflective" self-consciousness is not only an awareness of oneself at any given time, and of past and future phases of oneself, but also an awareness of one's own life as a whole, including being able to grasp, at least inchoately, the moral significance of one's own death. Therefore the capacity for principled authenticity also entails the possession of a concept of oneself *as* a conscious, intentional, caring, rational animal, and as a real person who is capable of free agency.[461]

That brings me, finally, to my second mid-sized point of substantive disagreement with Frankfurt. This concerns his notion of a "wanton." Here I have two worries.

First, I think that it is false that all non-human minded animals are wantons. On the contrary, in view of strong evidence from cognitive ethology, it is clear that at least some actual non-human minded animals—and in particular, Great apes and arguably also dolphins, perhaps parrots as well—are, at the very least, real non-human persons in the Frankfurtian sense. Hence, at the very least, they are metaphysically and morally equivalent to normal toddlers and other young children. On the opposing side, there seems to be some empirical neurobiological evidence in support of the claim that Great apes are not capable of non-instrumental, altruistic, non-hedonistic, and non-consequentialist thinking, feeling, and action.[462] But other primatologists would disagree with those skeptical primatologists, and I am on their side. It is one thing not to be capable of *sustained* non-instrumental, non-egoistic, non-hedonistic, and non-consequentialist thinking, feeling, and action. And this is perfectly consistent with being capable of brief moments of non-instrumental, non-egoistic, non-hedonistic, and non-consequentialist thinking, feeling, and action. Therefore, it is also perfectly consistent with the possession of all the basic innate capacities in such a way that some of them are not fully online. But it is categorically a different thing to lack those basic capacities altogether. My critical proposal is that those skeptical primatologists who have claimed the latter, have done so via a fallacious direct inference from the former. So, assuming that my critical proposal is correct, the fact that Great apes are not capable of *sustained* non-instrumental, non-egoistic, non-hedonistic, and non-consequentialist thinking, feeling, and action, does not in any way undermine their fully online Humean or lower-level (2D) rationality, and therefore it does not in any way undermine the Frankfurtian real personhood of Great apes, or their dignity. Correspondingly, we do not think that toddlers and other young children lack real personhood or dignity just because they often or even usually feel, choose and act, well, *childishly*.

Second, while it is true that in some *extended* sense of the term there can be "rational wantons," it is false that they are *not* real persons. More precisely, on my view, so-called "rational wantons" are in fact real persons who do indeed have an online capacity for second-order volitions, hence for caring about their caring, but for reasons of their own, in some context or contexts, they simply refuse to manifest or realize this capacity. When thought through carefully, we can see that it is inconceivable that a 2D rational animal could possess either non-autonomous, lower-level (Humean) or autonomous, higher-level (Kantian) rationality, and thus be capable of norm-guided logical and practical reasoning according to instrumental or non-instrumental principles, and yet also be *unable* to care about the constitution of its own will. This is because, for a creature to be inherently constrained by logical and practical norms in its reasoning, is necessarily *also* for it *to be able to care* about the difference between its freedom and its unfreedom. Otherwise, these norms would have no role to play in the conscious, intentional life of the creature: norms have to matter *to* and *fo*r the creature, and this cannot be if the creature cannot represent the difference between its own freedom and unfreedom.

Still, even accepting that point, it remains true, as a matter of fact, that some rational human animals or real human persons simply *refuse to care* about their caring, in some context or contexts, even though they are *able* to do so. Take, for example, Meursault in Camus's *The Stranger* as a vivid fictional instance of what I think is a real-world personality type. Meursault murders an Arab seemingly for no reason other than that he just feels like doing it at that moment; and he does it without any compunction whatsoever. Given Camus's descriptions and the rest of the narrative, Meursault is clearly as close as one could ever get, inside fiction or outside in the real world, to being a so-called "rational wanton" in Frankfurt's sense. Nevertheless, Meursault is also, clearly, deeply morally responsible for murdering the Arab.

And this is so, even despite its being true that, because Meursault is a murderer (*meurtrier*) who, seemingly without any premeditation whatsoever, leaps (*sauter*) into the evil act, it is quite hard to judge whether his act is a special sub-kind of *banal* evil, or *near-satanic* evil. In any case, from Camus's descriptions, it is clear not that Meursault *cannot* care about his caring—but rather that, although Mersault *can* care about his own caring, for whatever reason, in this context, he *refuses* to.

Therefore, a creature can be criticized and evaluated for her logical and practical deliberations, decisions, and choices only if it *can* matter to her which choices she makes. Interestingly, although Frankfurt initially explicitly holds that there can be rational wantons,[463] he later corrects himself by asserting that rationality entails personhood, which for him entails non-wantonhood.[464] So in that respect, as in others, he eventually came over to my contemporary Kantian "team." In any case, my view, by contrast, is that so-called "rational wantons" are indeed real persons, and therefore are *not constitutively wanton*, precisely because rational free agency entails at least a capacity for second-order volitions, even if, for some special reason or reasons intentionally adopted by a free agent in some context, that capacity is suppressed or non-operative in that context.

6.4 Conclusion

All things considered, my two contemporary Kantian extensions of Frankfurt's theory of persons are of central importance to the views I have been developing in this book, although they capture only mid-sized substantive disagreements with Franksfurt's overall account.

Let us assume, as I have just argued, that some non-human animals and all normal human toddlers and other young children are real persons; that both persons_f and persons_k are real persons, since an innate capacity to care about one's own or another's caring is minimally sufficient for real personhood; and that so-called "rational wantons" are not constitutively wanton, and are also real persons. Then it follows that the innate capacity for (2D) rationality in the higher-level or Kantian sense, including the innate capacity for deep moral or non-moral responsibility, and for principled authenticity—an innate capacity that appears to be fully possessed, in our actual world, only by adult, mature, sane humans—is *not* a necessary condition of real personhood, although of course it remains a sufficient condition. On the contrary, the universally necessary and minimally sufficient condition of real personhood is just real personhood in the Frankfurtian sense: the possession of an innate capacity for second-order volitions by an *S*-type animal, whether human or non-human, whether adult, adolescent, or young child, and whether or not this is accompanied by innate capacities for categorically norm-guided logical or practical reasoning, for self-reflective self-consciousness, for deep moral or non-moral responsibility, and for principled authenticity.

Furthermore, these considerations directly imply the existence of a distinct class of real persons between the class of *S*-type animals who are non-persons on the one hand, and the class of autonomous, higher-level, or Kantian (senior rank) real persons on the other, namely, the class of non-autonomous, lower-level, or Frankfurtian (junior rank) real persons.[465]

Or otherthwise put, and to end the chapter with two word-bites:

First, real personhood and being able to care about one's own or another's caring *are one and the same thing*.

Second, all normal human toddlers, other normal young children, and some non-human animals, for example, Great apes and arguably also dolphins, perhaps parrots as well, *are real persons too*.

Chapter 7

MINDED ANIMALISM II: FROM PARFIT TO REAL PERSONAL IDENTITY

> Like my cat, I often simply do what I want to do. I am then not using an ability that only persons have. We know that there are reasons for acting, and that some reasons are better or stronger than others. One of the main subjects of this book is a set of questions about what we have reason to do. I shall discuss several theories. Some of these are moral theories, others are theories about rationality.
>
> We are particular people. I have my life to live, you have yours. What do these facts involve? What makes me the same person throughout my life, and a different person from you? And what is the importance of these facts? What is the importance of the unity of each life, and of the distinction between different lives, and different persons? These questions are the other main subject of this book.
>
> My two subjects, reasons and persons, have close connections.[466]

7.0. INTRODUCTION

Derek Parfit's *Reasons and Persons*, which first appeared in 1984, is a widely influential classic of philosophical analysis in the methodological mode of the Standard Picture.[467] This big and brilliant book is about what we have reason to do, about the nature of persons and their personal identity over time, and also about the intimate and even necessary connections between these notions, especially insofar as they relate directly to morality and rationality. All the basic conclusions of *Reasons and Persons* are either explicitly reasserted or implicitly sustained by Parfit's second, last, and even bigger, and also brilliant, book, in two volumes, *On What Matters*, published in 2011, which fuses Kantian ethics to his earlier doctrines. The juxtaposition of *Reasons and Persons* and *On What Matters*, in turn, yields an important dilemma for Parfit's overall view, and also provides a smooth theoretical segue to my account of the nature of real persons and real personal identity.

In *Reasons and Persons*, Parfit wants to defend two basic theses.

First, we are mostly wrong in our commonsense beliefs about reasons and persons, and changing our beliefs about persons is essential to changing our beliefs about reasons:

> I believe that most of us have false beliefs about our own nature, and our identity over time, and that, when we see the truth, we ought to change some of our beliefs

> about what we have reason to do. We ought to revise our moral theories, and our beliefs about rationality.[468]

In other words, from my contemporary Kantian point of view, *Reasons and Persons* would be most illuminatingly entitled *Reasons Without Any Real Persons*.

Second, we at least sometimes, or possibly even always, have reason to do—that is, it is at least sometimes, or possibly even always rational, to do—what is not in our own self-interest:

> [The Self-Interest Theory of Rationality] claims that, whatever the cost to others, a rational agent *must* be biased in his own favour, even if, in a cool hour, he neither has nor wants to have this bias. Is this claim plausible? Is this bias uniquely or supremely rational? *This is the central question*…. No. I claim that, compared with the bias in one's own favour, there are several other desires that are no less rational. One example is a desire to act in the interests of other people. It can be rational to fulfil this desire, even when one knows that one's act is against one's own self-interest. Other examples are certain kinds of desire for achievement. A creator may want his creations to be as good as possible. A scientist, or philosopher, may want to make some fundamental discovery, or intellectual advance. I claim that these and other desires are no less rational than the bias in one's own favour. If one of these is someone's strongest desire, all things considered, it would be rational for him to cause it to be fulfilled, even if this person knows that his act is against his own self-interest.[469]

In other words, Parfit's second basic thesis is *that non-self-interested action is rational*. Since moral theories are almost always based on claims about rationality, and since surely it is the case that something's being the all-things-considered rational thing to do is at least sufficient, other things being equal, for its being morally obligatory, it then follows that we should accept a non-self-interested moral theory. A non-self-interested moral theory, in turn, holds that at least sometimes we ought to do what is not exclusively in our own self-interest, but instead what is exclusively in the interests of other people, even to the point of violating our own self-interest, hence something that is altruistic, or at least non-egoistic, and purely for the sake of Beauty, Truth, or some other "transcendental" value—hence, presumably, something that is not only non-egoistic, but also non-hedonistic and non-act-consequentialistic, and so-on.

Chapters 14 and 15 of *Reasons and Persons* offer two more important sub-theses. The momentous sub-thesis of chapter 14 is that if we accept Parfit's view about personal identity, then the egoistic or Self-Interest Theory of Rationality is false:

> On the unrevised or *Classical [Self-interest] Theory*, it is irrational for anyone to do what he believes will be worse for him. On the Revised Self-interest Theory, this claim may be abandoned. If it is not irrational to care less about some parts of one's future, it may not be irrational to do what one believes will be worse for oneself. It may not be irrational to act, knowingly, against one's self-interest.[470]

Correspondingly, the momentous sub-thesis of chapter 15 is that if we accept Parfit's view about personal identity, then we ought to make some radical changes to our views about morality and rationality, and this may significantly change our lives:

> We ought to be Reductionists. If this is a change of view, it supports several changes in our beliefs about both rationality and morality…. The effect on our emotions may be different for different people…. I find the truth liberating, and consoling. It makes me less concerned about my own future, and my death, and more concerned about others. I welcome this widening in my concern.[471]

Here are some specific examples of Parfit's proposed radical changes to our views:

> If we move from the Non-Reductionist to the Reductionist View, it becomes more plausible to claim that there is less scope for compensation within the same life. Thus it is more plausible to claim that great burdens imposed upon a child cannot be compensated, or fully compensated, by somewhat greater benefits in this child's adult life. When we thus extend distributive principles so that they cover, both whole lives, and weakly connected parts of the same life, this makes these principles more important. This is a move beyond the Utilitarian View.
>
> If we cease to believe that persons are separately existing entities, and come to believe that the unity of a life involves no more than the various relations between the experiences in this life, it becomes more plausible to be concerned about the quality of experiences, and less concerned about whose experiences they are…. The impersonality of Utilitarianism is therefore less implausible than most of us believe.
>
> If we become Reductionists, we can plausibly claim that a fertilized ovum is not a human being, and that it becomes a human being only gradually during pregnancy. This supports the claim that abortion is not wrong in the first few weeks, and that it only gradually becomes wrong.
>
> Some writers claim that only the deep further fact [about persons] carries with it desert, and that, since there is no such fact, we cannot deserve to be punished for past crimes…. I then argued for the general claim that, if the [R-relation] connections are weaker between a criminal now and himself at the time of this crime, he deserves less punishment. Similar claims applied to commitment.[472]

Now if one takes fully onboard, as I do, both Kantian and Existentialist ethical insights about the self-legislating 2D rationality of autonomous, higher-level, or Kantian intentional agency, and also about the categorically normative necessity of our having capacities to have wholehearted higher-order emotional concerns for the sake of the dignity of real persons and the Categorical Imperative, innately specified within us, then Parfit's second central thesis is self-evident: *Of course*, it is true that non-egoistic, non-hedonistic, or non-act-consequentialistic action and emotional concern is not only rational, but also really possible, even for "human, all too human" creatures like us. Even toddlers and other young children—although not Parfit's cat—are *sometimes* capable of it. On contemporary Kantian and Existentialist grounds alone, then, I believe along with Parfit that we rationally should accept a non-egoistic or non-self-interested, non-hedonistic, and non-act-consequentialistic moral theory of some sort. So I also believe along with Parfit that we rationally should accept that non-consequentialism is true.[473] Non-*act*-consequentialism, that is. Rule-consequentialism, by

contrast, as Parfit ably argues, ultimately collapses into Kantian moral theory. Correspondingly, part two of *On What Matters* is all about how the most plausible versions of rule-consequentialism and contractualism ultimately converge on a Kantian theory of moral principles and reasons; and in this regard I am also fully in agreement with Parfit.

At the same time, however, since my version of contemporary Kantian non-act-consequentialism follows from very different premises than Parfit's, I am not committed to, and indeed I have serious doubts about, Parfit's radical conclusions on distributive justice within individual lives,[474] on the moral importance of the quality of experiences as detached from the subject of experiences,[475] and on the morality of punishment,[476] and also at best a highly qualified agreement with his conclusions about the morality of abortion.[477]

In any case, where I think that contemporary Kantians should most sharply disagree with Parfit is about his first basic thesis—namely, that in order to demonstrate the rational compellingness of non-act-consequentialism, we must hold a radically non-commonsensical and Reductionist conception of the nature of persons and personal identity over time. This thesis is false. On the contrary, as I have been arguing, we can get to the rational compellingness of non-act-consequentialism if and only if we adopt a *real metaphysical* conception of real personhood and real persons, and also a correspondingly *real metaphysics* of real personal identity, that

> (i) closely conform to a version of contemporary Kantian Ethics that is appropriately combined with Existentialist ideas,
> (ii) closely conform to the mind-body thesis that necessarily all animals with a capacity for consciousness like ours are essentially embodied, and
> (iii) closely conform to the Natural Libertarianism theory of free agency that I worked out in chapters 1 to 5.

Or more briefly put for the present purposes of argument: we can get to the rational compellingness of any specifically contemporary Kantian version of non-act-consequentialism only if we adopt Minded Animalism. So to that extent, Parfit is fundamentally mistaken in *Reasons and Persons* and *On What Matters* alike, and the Partfitian is thereby faced with the following dialectical dilemma:

> (i) s/he must *either* give up Parfit's earlier reductive theory of persons and personal identity in *Reasons and Persons* (and retain his later Kantianism, plus a methodological switch from the Standard Picture to real metaphysics), *or else*
> (ii) s/he must give up Parfit's later Kantianism (and retain his earlier reductive theory of persons and personal identity, alongside the philosophical methodology of the Standard Picture).

In the rest of this final chapter, I will argue directly for Minded Animalism by using Parfit's theory of personal identity as a constructive critical foil. Obviously, it is a direct implication of this conclusion that the Parfitian should face up to the dilemma I just formulated, by *seizing the first horn of the dilemma*, that is, by

(i) giving up Parfit's earlier reductive theory of persons and personal identity,
(ii) going over to Minded Animalism and the philosophical methodology of real metaphysics, and
(iii) retaining Parfit's later Kantianism.

7.1 PARFIT'S THEORY: SIX BASIC CLAIMS

Parfit's theory of persons and personal identity can be expressed as the conjunction of six core claims.

First, he makes a claim about the nature of a person:

> To be a person, a being must be self-conscious, aware of its identity and continued existence over time.[478]

This is formulated as a necessary condition, but the rest of Parfit's discussion makes it clear that he also takes it to be a sufficient condition. In other words, Parfit's conception of personhood closely tracks the classical Lockean thesis that a person is essentially a continuing self-conscious subject of experiences.[479]

Second, he makes a claim about the criterion of personal identity over time:

> Our identity over time just involves (a) relation R—psychological connectedness and/or psychological continuity—with the right kind of cause, provided (b) that this relation does not take a "branching" form, holding between one person and two different future people.[480]

In other words, Parfit adopts a modified version of the psychological criterion of diachronic personal identity proposed by Locke, namely one that focuses on psychological continuity of all sorts (not just on continuity of memory, as in the conventional version of Locke's account[481]) and explicitly excludes person-fission cases.

Third and perhaps most importantly, he (Parfit that is, not Locke) makes a claim about Reductionism about persons:

> We are not separately existing entities, apart from our brains and bodies, and various interrelated physical and mental events. Our existence just involves the existence of our brains and bodies, and the doing of our deeds, and the thinking of our thoughts, and the occurrence of certain other physical and mental events.[482]

In other words, for Parfit persons are either identical with or strongly supervenient on brains and bodies, and on the various interrelated physical and mental events associated with those brains and bodies. It is somewhat unclear from his description of Reductionism about persons, whether the strong supervenience relation is supposed to be *logical* supervenience, so that persons are "nothing but" or "nothing over and above" their brains and bodies, or *natural* or *nomological* supervenience, which as we have already seen in chapters 1-2 above, is consistent with non-reductive physicalism about mental properties and events.[483] But in either case,

Parfit's approach to *persons* remains fully reductive, even if his corresponding approach to the mind-body problem would count, strictly speaking, as a version of non-reductive physicalism.

Fourth, he makes a claim about the determinacy of (our concept of) personal identity:

> It is not true that our identity is always determinate. I can always ask 'Am I about to die?' But it is not true that, in every case, this question must have an answer, which must be either Yes or No. In some cases this would be an empty question.[484]

In other words, according to Parfit, all our thought and talk about personal identity has really to do with our ways of thinking and talking about things that are not persons, and these ways of thinking and talking are sometimes truth-valueless.

Fifth, he makes a claim about what fundamentally needs to be explained about the fact of personal identity:

> There are two unities to be explained: the unity of consciousness at any time, and the unity of a whole life…. These unities must be explained by describing the relations between these many experiences, and their relations to this person's brain.[485]

In other words, for Parfit the metaphysical bases of personal identity are two distinct types of unity—the unity of consciousness, and the unity of a whole conscious life that includes the unity of consciousness as a necessary condition—and once we have explained these two unities, then we have explained personal identity.

Sixth and finally, Parfit makes a claim about what really matters with regard to persons:

> [B]eing destroyed and Replicated is about as good as ordinary survival.[486]

> Personal identity is not what matters. What fundamentally matters is relation R, with any cause. This relation is what matters even when, as in a case where one person is R-related to two other people, Relation R does not provide personal identity. Two other relations may have some slight importance: physical continuity, and physical similarity.[487]

In other words, what really matters with regard to persons—who are construed by Parfit as being, essentially, continuing self-conscious subjects of experiences—is not personal identity at all. Rather what really matters is just psychological continuity and the physical strong supervenience base of persons.

Unfortunately for Parfitians, all six of Parfit's core claims are false. In the next two sections I will spell out an alternative theory of personal identity—Minded Animalism—and also indicate more precisely where Parfit's reductive account has gone wrong.

7.2 Against and Beyond Parfit 1: Two Reasons, and the Minded Animalist Criterion of Personal Identity

In one sense, my disagreement with Parfit's theory of persons and personal identity is simple: Parfit's theory about personhood and personal identity is just plain *false*, precisely

because it does not conform *either* to our prima facie, pre-theoretic, or commonsensical, rational intuitions about them, *or* to our phenomenologically self-evident claims about them, *or* to our authoritative real metaphysical and moral rational intuitions about them. In short, Parfit's theory does not tell us about *real human persons* at all. So as I noted above, his first book should have been entitled *Reasons Without Any Real Persons.*

It is obvious enough, then, that his theory does not conform to our prima facie, pre-theoretic, or commonsensical, rational intuitions about real personhood and real personal identity, *since he explicitly says this himself*. But Parfit's theory also does not conform either to our phenomenologically self-evident claims about them or to our authoritative real metaphysical and moral rational intuitions, precisely because there is at least one other theory that

> (i) smoothly conforms to all our prima facie, pre-theoretic, or commonsensical rational intuitions, *and* to our phenomenologically self-evident claims, *and also* to our authoritative real metaphysical and moral rational intuitions about real personhood and real personal identity,
> (ii) smoothly conforms to whatever relevant data and facts the formal and natural sciences can contribute to our understanding of real personhood and real personal identity,
> (iii) is internally consistent, and also
> (iv) entails the denials of each of Parfit's six core claims.

This theory is Minded Animalism. To keep things orderly, I will address each of Parfit's core claims in turn.

First, Parfit is wrong about the nature of a person. This is because a person in the relevant sense is just what I said it is in chapter 6, namely, a *real person*: a conscious, self-conscious, intentional, caring, 2D rational, *S*-type animal capable of free agency, whether possessed of non-autonomous, lower-level, or Frankfurtian personhood (junior rank personhood) and innate capacities for Humean instrumental 2D rationality, and for deeply free volition, or also possessed of autonomous, higher-level, or Kantian personhood (senior rank personhood) and innate capacities for non-instrumental 2D rationality, for deep freedom and deep (non-)moral responsibility, and for the achievement of principled authenticity at least partially or to some degree.

Second, as a consequence of my real-metaphysics-based analysis of personhood as real personhood, together with the further fact that real personal identity is an essential fact about each and all the integral parts of the whole single, minimally coherent, intrinsically spatiotemporal, complete, finite, and unique, dominantly free, essentially embodied life of a free agent, it follows that Parfit is also wrong about the criterion of diachronic personal identity—our persistence throughout a whole life over time. According to Minded Animalism, a real person X is one and the same thing as Y if and only if

> (i) X is a minded animal and Y is a minded animal,
> (ii) X is a *a real person* in the sense defined in chapter 6, whether a Frankfurtian non-autonomous real person or a Kantian autonomous real person, and so is Y,

> (iii) X and Y stand in relations of intrinsically spatiotemporal, essentially embodied, biophenomenological continuity and intrinsically spatiotemporal biological/neurobiological continuity with each other, and
> (iv) X has a freedom-dominated life, which is to say that enough of the intrinsically spatiotemporal, essentially embodied, dual biophenomenological and biological/neurobiological events constituting the continuous life of X are also deeply free, flow from her ultimate sourcehood, and up to her, and so does Y.

Thus the Minded Animalism theory of personal identity entails an essentially *triadic*, three-factor criterion of real personal identity over time, in that it involves a biophenomenological condition, a biological/neurobiological condition, and a freedom-dominance condition that are individually necessary, and individually insufficient, but also jointly sufficient for real personal identity. More precisely, according to this three-factor criterion of personal identity, a real person *X* is one and the same thing as *Y* if and only if

> (i) X is a minded animal and Y is a minded animal,
> (ii) X is a real person in the sense defined in chapter 6, whether a Frankfurtian non-autonomous real person or a Kantian autonomous real person, and so is Y,
> (iiia) X is intrinsically spatiotemporally, essentially embodied-ly,[488] biophenomenologically continuous with Y, which is to say that X has enough of the same consciousness-based online psychological capacities and enough of the same conscious, intentional, caring, 2D rational contents as Y,
> (iiib) X is intrinsically spatiotemporally, essentially embodied-ly, biologically and neurobiologically continuous with Y, which is to say that X has enough of the same necessary and complete biological/neurobiological embodiment as Y, and
> (iiic) X has a freedom-dominated life, which is to say that enough of the essentially embodied, dual biophenomenological and biological/neurobiological events constituting the continuous life of X are also deeply free, flow from her ultimate sourcehood, and up to her, and so does Y.

For terminological convenience, I will call this triadic criterion *The Minded Animalist Criterion of Personal Identity*.

It should be particularly noted in this connection that the biophenomenological condition (iiia) fully presupposes the distinction I made in section 6.2 between

> (i) pre-reflective, non-self-conscious, or first-order consciousness, and
> (ii) self-reflective or self-conscious consciousness.

This is directly relevant to an important thought-experiment, developed by Bernard Williams, intended to show that I am identical with my body. In a nutshell, Williams argues, quite plausibly, that I would fear being tortured tomorrow even if I underwent a brain-washing and consciousness-replacement process between now and then.[489] Where he goes wrong, however, is in concluding from this thought-experiment that I am my body. That is simply a non sequitur. More precisely, from the standpoint of Minded Animalism and The Minded Animalist Criterion of Personal Identity, there are two critical things to say about Williams's thought-experiment.

First, it is perfectly possible to have significant moral concern, and even a special selfish concern, for creatures who are in various ways similar to me, even if they are not identical to

me. I will come back to this point below in my critical discussion of Parfit's account of personal identity.

Second, and most importantly in the present context, the rationally intuitive force of Williams's thought-experiment depends almost entirely on the implicit, unarticulated assumption that I will retain a pre-reflective, non-self-conscious, or first-order conscious biophenomenological continuity with the human animal that will feel pain when it is tortured tomorrow, even if I do not retain a *self-reflective or self-conscious* biophenomenological continuity with the human animal that will feel pain when it is tortured tomorrow. If, for example, I were reliably informed that either my own corpse or a decerebrated ("brainless") counterpart of my body were going to be tortured tomorrow, then obviously I would not fear that. It is only if the biophenomenological continuity of my essentially embodied consciousness is preserved, at least at the pre-reflective, non-self-conscious, or first-order conscious level, that I would fear my body's being tortured tomorrow. Even more precisely, however, what I would fear is my own *body-based suffering* tomorrow.[490] Of course, I am not intending to imply that either my own corpse or a brainless counterpart of my own body could feel bodily pain—on the contrary, it is obvious that both body-based suffering and also mere bodily pain occur *only in minded living organisms*.

It should also be particularly noticed in this connection that the third factor in The Minded Animalist Criterion of Personal Identity now provides us with a fourth and final argument[491] for Local Incompatibilism with Respect to Natural Mechanism, as follows.

> Argument 4: The Personal Identity Argument for Local Incompatibilism with Respect to Natural Mechanism
>
> (1) If Natural Mechanism is true, then both the existence and the specific character of whatever I'm apparently choosing or doing at any time aren't up to me. (From argument 1 above in section 4.5, step (3), and argument 2 above in section 4.5, step (3).)
> (2) If both the existence and the specific character of whatever I'm apparently choosing or doing at any time aren't up to me, then none of the essentially embodied, biophenomenologically and biologically/neurobiologically continuous events that are causal-dynamically associated with my life are really my own. (From (1), and the phenomenologically self-evident, or basic authoritative rational-intuitive notions, of deep freedom and deep (non-)moral responsibility, together with The Minded Animalist Criterion of Personal Identity.)
> (3) So if Natural Mechanism is true, then I lack real personal identity. (From (1), (2), and the freedom-dominance condition on real personal identity.)
> (4) So if Natural Mechanism is true, then I am not a free agent. (From (3) and the ownership condition on deep freedom.)

And here is one last thing to be particularly noticed about The Minded Animalist Criterion of Personal Identity, before moving on. The Minded Animalist Criterion fully allows for the fact that I, who am now an actualized real person, can be literally identical with something—for example, my third-trimester conscious fetus—that was not yet an actualized real person. Indeed, on my view, the beginning of a real person's life, or its "neo-personhood," is when a

given *S*-type animal *A* manifests the psychological capacity for consciousness and the following counterfactual is also true of *A*:

> If *A were* to continue the natural course of its neurobiological and psychological development, then *A would* become an actualized real person.

This conception of neo-personhood has important implications for the morality of abortion and infanticide.[492] But in the present context I want to emphasize its important real-metaphysical implications for the doctrine of real personal identity that I am proposing. These real-metaphysical implications have to do with the nature of the connections between certain special kinds of temporal parts and temporal wholes.

"Metonymy*"* is a relation whereby a proper part of a whole stands for the whole of which it is a proper part. According to the doctrine of real personal identity that I am proposing, real personal identity is a special mereological (that is, part-whole) relation of *metonymous identity in spacetime*, whereby

> (i) a proper spatiotemporal part of an immanently-structured unified spatiotemporal whole is strictly identified with the immanently-structured unified spatiotemporal whole of which it is a proper part,
> (ii) that immanently-structured unified spatiotemporal whole is strictly identified with itself,
> (iii) that immanently-structured unified spatiotemporal whole is also strictly identified with that proper spatiotemporal part, and,
> (iv) each and every spatiotemporal part (whether proper or improper—that is, the complete collection of proper parts, bound together by the unifying immanent structure of the whole) of that immanently-structured unified spatiotemporal whole meets conditions (i), (ii), and (iii).

By "classical identity," I mean the relation of necessary numerical identity, including the properties of symmetry, transitivity, and reflexivity, plus satisfaction of Leibniz's Laws for all non-modal, non-normative, and more generally non-intensional properties, and intersubstitutivity *salve veritate*. As I am understanding it, then, the real personal identity relation is symmetrical, transitive, and reflexive, but these familiar formal features of classical identity are applied *over* the mereological metonymy relation between proper finite spatiotemporal parts and immanently-structured unified finite spatiotemporal wholes. These, in turn, are not numerically (aka "token-token") identical. Leibniz's Laws fail for them. They do not share all their non-modal, non-normative, and more generally non-intensional properties properties in common And you cannot freely substitute one for the other, without changing anything. On the contrary: if you tried to intersubstitute proper finite spatiotemporal parts and their immanently-structured unified finite spatiotemporal wholes, then everything would go all pear-shaped. But they are, nevertheless, *literally the very same*. That makes the real personal identity relation a conservatively non-classical but still modally strict (that is, non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" necessary) identity relation, just like most other accounts of personal identity, but now in a special mereological and metonymous way.

Moreover, the special mereological relation of metonymous identity is not at all unique to personal identity. For example, every successive moment of a single day is literally the very same day, and that very day is metonymously present in each and all of its successive moments. Every successive note of a single piece of music is literally the very same piece of music, and that very piece of music is metonymously present in each and all of its successive notes. Every successive play in a single game is literally the very same game, and that very game is metonymously present in each and all its successive plays. And the same holds, mutatis mutandis, for all proper parts p_1, p_2, $p_3 \ldots p_n$ of every irreversible, goal-directed spatiotemporal process *P*. So too, then, every successive stage of my complete, finite, and unique life is literally the very same life, and that very same life is metonymously present in each and all of its successive stages.

People often say, with some truth, and drawing at least implicitly on a famous Kierkegaardian metaphor in *Stages on Life's Way*, that a person's life is a journey. Yet it is much more accurate, although somewhat less catchy, to say that each real person's life is a holistic, essentially embodied, biophenomenologically, biologically-neurobiologically, and volitionally irreversible, goal-directed spatiotemporal process, continuously divisible into a plurality of distinct spatially situated and temporally irreversible ordered stages, each bearing the personal identity relation to each and all of the other stages and thus to the whole life itself. In other words, according to Minded Animalism, the metaphysics of real personal identity is a mereological, natural-teleological, and process-based dynamicist real metaphysics—as it were, *macroscopic Whiteheadianism*[493]—and not primarily a real metaphysics of material substances, although it does also include material substances as derivative or secondary facts about self-organizing complex thermodynamic systems. Each preliminary stage prior to the achievement of actual real personhood marks a proper spatiotemporal part of a complete, finite, unique real personal life that is essentially a unified spatiotemporal process that naturally terminates and is teleologically completed only at death.

Thus the minded *S*-type animal or living organism, is, quite literally, identically the same real person from the very beginning of that life when it is a neo-person (for normal humans, roughly 25-32 weeks after conception or fertilization), and also at every distinct stage through the point or threshold at which it reaches actual non-autonomous, lower-level, or Frankfurtian real personhood (for normal humans, roughly one year of age), and beyond, until it later reaches actual autonomous, higher-level, or Kantian real personhood (for normal humans, roughly eighteen years of age). Then, still, quite literally, identically the same real person, it passes again beyond that stage, all the way to its death—which, for normal humans living in Canada or the USA, other things being equal,[494] amounts to a span of roughly four score, or 80, years.

So, assuming that real personal identity is a mereological relation of metonymous identity between proper temporal parts and immanently-structured unified spatiotemporal wholes that are unique, finite, and complete irreversible, goal-directed life-processes running from births to deaths, then while it is true that real human persons are always and everywhere essentially real human persons—hence a real human person cannot be self-identical with anything other than a real human person—is also a mistake to hold that self-identical real human persons are always and everywhere *actualized* real human persons. They can also be what I call *strongly potential* real human persons.[495] Correspondingly, the difference between the actuality and the strong potentiality, respectively, of actualized real persons and the strongly potential earlier versions of them with which they are strictly identical, is one of those modal and intensional

differences to which classical identity and its non-classical conservative extensions are systematically insensitive.

As I pointed out above, before we humans achieve fully-constituted real personhood at approximately one year of age, we are "neo-persons," where neo-persons are all and only those animals or living organisms that have manifested the capacity for consciousness, and are not yet actualized real persons, but will become actualized real persons in the natural course of their later neurobiological and psychological development, other things being equal. Neo-persons are part-to-whole identical with the actualized persons they later become. Or more precisely, neo-persons are identical *with what they will be*, namely actualized real persons, provided that certain further *ceteris paribus* conditions are satisfied. But classical identity and its conservative extensions requires both of its terms to have the same kind of existence.

Formulated in linguistic-semantic terms, my view is that the term "real person" is not a *substance sortal term* that applies to one and only one substance, where this substance is classically understood to be a unchanging bearer of changing attributes whose nature is determined by intrinsic non-relational properties, and which therefore is able to exist (at least as a matter of logical, conceptual, analytic, or "weak metaphysical" possibility) even if everything else fails to exist—thereby having what David Lewis evocatively called a "lonely existence." Nor, according to my view, is the term "real person" a *phase sortal term* that applies to such a substance for only certain parts of its total spacetime career.[496] Instead, by contrast to both of these, the interpreted noun phrase "real person" is what I will call a *process sortal term*. More specifically, "real person" applies to each and all of the distinct spatiotemporal proper parts, and also to the unified spatiotemporal whole composed of these, of a far-from-equilibrium, asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned, conscious, intentional, caring, 2D rational, deeply free and deeply (non-)morally responsible thermodynamic process, insofar as all those spatiotemporal parts are indeed continuously bound together into a single, multi-termed, immanently-structured unfied spatiotemporal whole, and each distinct proper part bears a reflexive, symmetrical, and transitive *literally-the-very-same-S* relation to each and all parts of that immanently-structured unified spatiotemporal whole, under that sortal *S*.

In this regard, then, the syntax and semantics of the term "real person" are highly similar to the syntax and semantics of these more familiar and ordinary process sortal count-nouns in English, listed alphabetically:

career
century
dance piece
day
decade
endeavor
era
game
hour
journey
millennium
minute
month

piece of music
project
semester
term
undertaking
war
week
year

and—perhaps not altogether surprisingly—

life
lifetime.

The Minded Animalist Criterion of Personal Identity can also be smoothly extended to real personal identity over space at a given time—that is, synchronic personal identity—by restricting conditions (iiia), (iiib), and (iiic) to a single time (whether a moment or a duration), which guarantees localized biophenomenological and biological/neurobiological continuity, and localized freedom-dominance. On this reformulation of The Minded Animalist Criterion of Personal Identity—with changes underlined—a person *X* is one and the same thing as *Y* at that same time (whether a moment or a duration) if and only if

> (i) X is a minded animal and Y is a minded animal,
> (ii) X is a real person in the sense defined in chapter 6, whether a Frankfurtian non-autonomous real person or a Kantian autonomous real person, and so is Y,
> (iiia) X is intrinsically spatiotemporally, essentially embodied-ly, biophenomenologically continuous with Y at that same time, which is to say that X has enough of the same consciousness-based online psychological capacities and enough of the same conscious intentional contents as Y at that same time,
> (iiib) X is intrinsically spatiotemporally, essentially embodied-ly, biologically and neurobiologically continuous with Y at that same time, which is to say that X has enough of the same necessary and complete biological/neurobiological embodiment as Y at that same time, and
> (iiic) X has a freedom-dominated life at that same time, which is to say that enough of the dual biophenomenological and biological/neurobiological events constituting the continuous life of X at that same time are also deeply free and up to her, and so does Y.

This version of The Minded Animalist Criterion of Personal Identity has a direct bearing on cases of dicephalus (or two-brained conjoined) twins, and ensures that they are distinct real persons by virtue of their distinct necessary and complete neurobiological embodiments, despite their sharing all or most of the same vital organs other than their brains.

A closely related but even more complex kind of case of conjoined twins is the real-world case of twins in the Philippines, born in 2002, who originally shared a single "fused" brain, although each possessed a distinct set of all their other vital organs, and were later surgically separated in 2004.[497] For clarity's sake, I will call twins of this sort *janus* conjoined twins[498] in

order to distinguish them from *dicephalus* conjoined twins. Prior to their surgical separation, the Filipino janus twins were said by neurosurgeons to share literally the same consciousness and literally the same thought-processes, and yet also have somewhat distinct personalities—one of them, for example, would spontaneously reach around and smack the other, and appeared to enjoy it, while the other did not appear to enjoy it. The surgical separation was "successful," in the sense that it produced two fully distinct 2 year-old human individuals and real human persons, Carl and Clarence Aguirre. The application of The Minded Animalist Criterion of Personal Identity to this subtly difficult but philosophically quite instructive case has three parts:

> (i) prior to surgical separation at the age of 2, the Filipino janus twins jointly constituted a single human individual[499] and a single real person—call him "Carl/Clarence Aguirre"—with a unique embodied consciousness and a unique complete neurobiological embodiment, but also with a *duplicated* set of vital organs (analogous to being born with a second set of teeth, or an extra nipple), and also a somewhat *dissociated* personality, in a way that is similar to neo-commissurotomy cases, including the strange case of the wife-smacking man,[500]
> (ii) the surgical separation of the Filipino janus twins produced two new real human persons, Carl A. and Clarence A, and
> (iii) the surgical separation of the Filipino janus twins therefore corpselessly killed Carl/Clarence Aguirre by destroying the neurobiological basis of his unique embodied consciousness.

In short, the surgical separation of the Filipino twins was an unusual case of *toddlericide*.

As such, there is a further and equally subtly difficult ethical question as to whether it was a morally permissible case of toddlericide, or not. It is unclear to me what the precise moral reasoning of the parents and doctors was, but let us suppose for the purposes of argument that on the basis of good medical evidence, they all reasonably believed that if surgical separation did not occur, then Carl/Clarence would die. Assuming that, then I think it was a case relevantly similar to the permissible Trolley Problem cases—in this case, killing 1 in order to save 2—and that it was therefore morally permissible to kill Carl/Clarence A. in order to create the two new distinct real persons, Carl A. and Clarence A. If, for the same moral reasons, the surgical separation had occurred during the neo-personhood of Carl/Clarence A., significantly prior to the advent of his Frankfurtian actualized real personhood (just for argument's sake, let us say that the advent of real personhood occurs at roughly at 6 months of age, and the operation occurred at 3 months of age), then it would also have been a morally permissible case of infanticide for essentially the same reasons.[501]

One further point in this connection. As I have said, according to The Minded Animalist Criterion of Personal Identity, since as far as we now know, fetal consciousness in normal human fetuses begins roughly 25-32 weeks after conception or fertilization, and since a real person's life begins when she enters the phase of strong potentiality for actualized real personhood, that is, when she first acquires consciousness, then it follows that I am identical with my third trimester fetus, a neo-person. But if, like the unfortunate Terry Schiavo, some time after I have achieved actualized real personhood I suffer a permanent loss of consciousness by the shut-down of most of my brain-functions and also a corresponding transition to a persistent vegetative state, then my real personal life will have ended in my death at precisely

the point of shut-down and transition, despite the fact that (most of) my living body still exists. Nevertheless, even if I escape the fate of Terry Schiavo, I might still at some point later in life—by the slings and arrows of accident, disease, or just aging—be reduced to the conscious and cognitive state of an infant, as for example, in the case of the brilliant philosopher Iris Murdoch during the later stages of her poignant succumbing to Alzheimer's disease.[502]

If so, then my real personal life will not have ended in my literal death (as in the Schiavo case), but instead (as in the Murdoch case) it will have ended in what I will call my *quasi-death*. My neurobiologically continuous successor after my quasi-death will be what, following Jeff McMahan's lead, I will call a *post-person*,[503] who is identically the same *S*-type animal or living organism as I am, although he will not share a real personal identity with me, since he fails the biophenomenological continuity criterion. As we will see below in my discussion of the basic thought-experiment case that Parfit calls Psychological Spectrum, post-persons have essentially the same metaphysical and ethical status as a biological-neurobiological continuant of me who retains some phenomenological continuity with me, but not enough to constitute his being a genuine biophenomenological continuant of me, which thereby triggers the termination of my real personal identity and the end of my life, that is, which thereby triggers my death. There is so much to say about the morality of the beginning and/or ending of a real person's life! As Dashiell Hammett's Sam Spade facetiously says in a slightly different but closely related connection, somebody ought to write a book about it.[504] In any case, the crucial point here is that there is a striking metaphysical and ethical asymmetry between

> (i) the beginning of a real person's life, which extends identically all the way back to one's neo-personhood and connects continuously with an *S*-type animal or living organism that has a strong potentiality for being a real person, and
> (ii) the end of a real person's life, for example, my death, even though there may still exist a biological-neurobiological continuant of me who also bears my proper name and has some minimal biophenomenological continuity with me.

The crucial difference that grounds this metaphysical and ethical asymmetry is the fact that it is true of my neo-person that he will *actually be me*, if he is allowed to go on living, whereas this is false of my post-person: he will *never actually be me*, no matter how long he goes on living. It is prospectively rather a sad thought, but still quite true, that my post-person will always be nothing but a "has-been" me.

7.3 Against and Beyond Parfit 2: Four More Reasons

Third, Parfit is wrong about Reductionism about persons. According to Minded Animalism:

> (i) it is false that a real person is either identical with or logically supervenient on a particular brain, body, and a series of interrelated physical and mental events, and
> (ii) it is also false that a real person is a separately existing entity over and above a particular brain, body, and a series of interrelated physical and mental events.

In sharp contrast to both of these, according to Minded Animalism, a real person is an inherently integrated and non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" necessarily indissoluble living psychophysical totality, an essentially embodied mind or minded animal, whose conscious will has a certain desire-based hierarchical constitution, and who is deeply free, with all that entails. Thus the failures of Reductionism (either downwards identity from the mental to the physical, or upwards logical supervenience of the mental on the physical), non-reductive physicalism (the strong supervenience of the mental on the physical, without reduction), and Dualism (the mutual modal independence of the mental and the physical) about real persons, alike, follow directly from the non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" necessary two-way *inter*dependence of a real person's fundamental mental and physical properties. And this, in turn, follows directly from The Essential Embodiment Theory of the mind-body relation.[505]

Fourth, Parfit is wrong that personal identity is not always determinate. On the contrary, real personal identity is always (in a real-metaphysical sense, at least, even if not epistemically) determinate, and The Minded Animalism Criterion of Personal Identity yields definite and authoritative rationally intuitive results for each one of the central thought-experiment cases he describes. I will now survey these central cases and briefly justify my philosophical rational intuitions about each of them.

***

Simple Teletransportation:

> At the beginning of my story, the Scanner destroys my brain and body. My blueprint is beamed to Mars, where another machine makes an organic *Replica* of me. My Replica thinks that he is me, and he seems to remember living my life up to the moment when I pressed the green button. In every other way, both physically and psychologically, we are exactly similar. If he returned to Earth, everyone would think that he was me.[506]

According to The Minded Animalist Criterion of Personal Identity, I clearly do *not* survive teletransportation, precisely because condition (iiib), that is, the biological/neurobiological continuity condition, is not satisfied. If my actual living body is destroyed, then I am destroyed, even if a replica of me with an exactly similar (but not literally the same) body and exactly similar (but not literally the same) psychology comes out of the teletransporter at the other end. No matter how positively disposed I might be towards my replica and his future career, he simply does not have *my own* essentially embodied conscious life. I do not survive what the Scanner did to me. I died on Earth, and my replica's life began on Mars. Since according to my view of the nature of properties, the orientable spatiotemporal properties of things are also intrinsic relational, or immanent structural, properties of those things, then it follows that I have one complete biological/neurobiological embodiment, and that my replica necessarily has a distinct complete biological/neurobiological embodiment. For these reasons, I would never let the original *Star Trek's* Scotty beam me up—or down, or sideways for that matter.

***

Callous Neuro-surgeon:

> I am the prisoner of some callous neuro-surgeon, who intends to disrupt my psychological continuity by tampering with my brain. I shall be conscious while he operates, and in pain. I therefore dread what is coming. The surgeon tells me that, while I am in pain, he will do several things. He will first activate some neurodes that will give me amnesia. I shall suddenly lose all of my memories of my life up to the start of my pain. Does this give me less reason to dread what is coming? Can I assume that, when the surgeon flips this switch, my pain will suddenly cease? Surely not. The pain will so occupy my mind that I would fail even to notice the loss of all these memories. The surgeon next tells me that, while I am still in pain, he will later flip another switch, that will cause me to believe that I am Napoleon, and will give me apparent memories of Napoleon's life. Can I assume that this will cause my pain to cease? The natural answer is again No. To support this answer, we can again suppose that my pain will prevent me from noticing anything. I shall not notice my coming to believe that I am Napoleon, and my acquiring a whole new set of apparent memories. When the surgeon flips the second switch, there will be no pain at all in what I am conscious of. The changes will be purely dispositional. It will only become true that, if my pain ceased, so that I could think, I would answer the question "Who are you?" with the name 'Napoleon'. Similarly, if my pain ceased, I would then start to have delusory apparent memories, such as those of reviewing the Imperial Guard, or of weeping with frustration at the catastrophe of 1812. If it is only such changes in my dispositions that will be brought about by the flipping of the second switch, I would have no reason to expect this to cause my pain to cease. The surgeon then tells me that, during my ordeal, he will later flip a third switch, that will change my character so that it becomes just like Napoleon's. Once again, I have no reason to expect the flipping of the switch to end my pain. It might at most bring some relief, if Napoleon's character, compared to mine, involved more fortitude.[507]

The Callous Neuro-surgeon case is obviously designed by Parfit to mimic Williams's torture-in-the-future case. According to The Minded Animalist Criterion of Personal Identity, I clearly *do* survive the surgery, precisely because condition (iiia), namely, the biophenomenological continuity condition, is satisfied at the level of my pre-reflective or first-order consciousness of experiencing bodily pain, even if it is not satisfied at the level of self-reflective or self-conscious consciousness. Now if my actual inner life, both at the level of pre-reflective or first-order consciousness and also at the level of self-reflective or self-conscious consciousness, were to have been permanently replaced by a pain-free psychic replica of Napoleon's inner life, then clearly according to The Minded Animalist Criterion of Personal Identity (iiia), I would not have survived the operation. But as the case is described in detail by Parfit, I definitely do continue to exist, painfully, in a pre-reflective or first-order conscious way, and therefore as ultimately as a *suffering* animal, throughout the operation, even despite my permanently acquiring the self-reflective or self-conscious aspects of Napoleon's life, and even despite my also permanently losing the relevant self-reflective or self-conscious aspects of my own life.

As I noted above, Williams's claim that I would fear being tortured in the future gains its intuitive force and plausibility entirely from the implicit or unarticulated assumption that I retain at least a pre-reflective, non-self-conscious, or first-order biophenomenological continuity, with my tortured body-to-be. On this construal of the situation, the biological/neurobiological basis of my essentially embodied consciousness continues to play the same functional role of detecting internal or external damage to the animal or living organism that I am now, and I also continue pre-reflectively or first-order consciously to "live" my living animal body, which then tomorrow becomes the vehicle of my body-based suffering.

Callous Neuro-surgeon also raises another very important and more general point about real human personal identity. I have so far assumed that when the neuro-surgeon flips one of the switches, the radical replacement of the memory-contents and character-traits of my conscious life by a replica of Napoleon's memory-contents and character-traits, is *permanent*. But what if this is only a temporary radical change of memory-contents and character-traits? And what about temporary insanity—as, for example, in the Ketema Ross case—temporary radical amnesia, temporary massive agnosia, temporary unconsciousness, deep sleep, and so-on?

Here we need to make a distinction between

> (i) *terminations* of my real personal identity (and, consequently, my *literal* death, even if it is only a quasi-death in which a post-person succeeds me and continues to exist), and
> (ii) *hibernations* of my real personal identity.

Although it is sometimes true that hibernations are also terminations, this is not always or perhaps even mostly so. A strict entailment from hibernations to terminations would hold if and only if the hibernation of my real personal identity is permanent. But in cases of temporary radical change of memory-contents or character-traits, temporary insanity, and so-on, then my real personal identity is only temporarily latent during that period, although I do not really personally literally die, and indeed I do really personally survive through that period in a dispositional sense. Otherwise put, my real personal identity remains in dispositional existence and resides in the temporarily latent and temporarily offline, but still undestroyed, psychological and volitional capacities that continue to exist by means of a necessary connection with the biological-neurobiological basis, or *natural matrix*, of my capacity for either pre-reflective/non-self-conscious/first-order or self-reflective/self-conscious consciousness.

This notion of temporary hibernations, temporary offline periods, or what I will call *latency phases* within one's real personal life, in turn explains one important way in which I can fail to be deeply (non-)morally responsible for something I only *apparently* choose or do. This is because the deeply free real person I am is temporarily not available during any latency phase in my ongoing real personal life, due to temporary radical change of memory-contents or character-traits, temporary insanity, one or more of my basic psychological capacities being temporarily offline, and so-on. During a latency phase, I am temporarily absent as a real person, and as we aptly colloquially say, I'm "not myself," "out of it," "a zombie," "zonked," or perhaps even, in extreme cases, "completely out of my head." In turn, and in either case, my merely apparent choices and acts, hence my "choices" and "acts" during a latency phase cannot be deeply free. Therefore, I am not deeply (non-)morally responsible for such "choices" and "acts"

of my own living organismic body during a latency phase, either. The phenomenology of merely apparent choices and acts, correspondingly, would exemplify *non-veridical psychological freedom*, not veridical freedom—see section 5.3 above, especially, again, the Ketema Ross case. But I do personally *survive* latency phases, so it is also colloquially correct to say that it is *I* who am "out of my head" and "not myself." Hence it is also at least colloquially correct to attribute scare-quoted "choices" and "actions" to the surviving hibernating and dispositionally deeply free me in a complete, finite, and unique life that exemplifies freedom-dominance. And this is so, even though those particular so-called "choices" and "acts" are not under my control, even though they are subjectively experienced as non-veridical psychological freedom, and even though the real person I am is not deeply morally responsible for them—again, see section 5.3 and the Ketema Ross case,

Therefore, on the one hand, (i) real personal *identity* according to The Minded Animalist Criterion of Personal Identity, and, on the other hand, what I will call (ii) real personal *manifestation*, that is, my real personal identity's being fully non-latent, are not necessarily equivalent. Although real personal manifestation strictly entails real personal identity, the converse does not hold: my real personal identity is consistent with the occurrence of latency phases within my real personhood.

As the point about non-veridical psychological freedom already indicates, the phenomenon of latency phases in real personal identity also directly raises a hard problem about those very phases in my essentially embodied biophenomenological and biological/neurobiological life that are *not* also phases in my life as a free agent. And by a natural extension, it also raises a hard problem about all those external facts in the world that directly affect who I am and what I choose or do, but over which I had and have little or no control whatsoever—the things that merely happen to me. The overarching hard problem here, as Nagel and Williams have correctly pointed out, and as I briefly discussed already in section 6.2 above, is precisely how to deal both metaphysically and morally with the phenomenon of *moral luck* insofar as it bears directly on my own real personal life, and more specifically how to deal with the brute fact of Problematic Episodes in my life, now fully generalized to the phenomenon of *non-moral* luck too: that is, taking *creative* responsibility for things in my life over which I had no control.

Now, I have said that in order to achieve principled authenticity, at least partially or to some some degree, I must not only choose and act wholeheartedly in accordance with my own personal projects, my moral principles, and with some absolute moral principles, but also I must sometimes take deep (non-)moral responsibility for things over which I had no control. This passionately stoic act of taking deep (non-)moral responsibility, in turn, can take the form of what I will call *appropriation*, which is

> either (i) my taking deep (non-)moral responsibility for some phases in my biophenomenological and biological/neurobiological life which were also latency phases in my real personal life in which certain things *merely happened* to me (I will call this "centrifugal" or inner→outer appropriation),
> or else (ii) my taking deep (non-)moral responsibility for some of the external facts in the world that directly affected who I am, and what I choose and do, but over which I had and have little or no control (I will call this "centripetal" or outer→inner appropriation),

in order to achieve what I called *authentic personal identity*. In other words, via centrifugal or centripetal appropriation, I can *resolve* at least some Problematic Episodes in my life by freely taking deep (non-)moral responsibility for things which merely happened to me or for some external facts over which I had no control, all of which were therefore facts for which I was not deeply (non-)morally responsible at that time.

If and when either centrifugal or centripetal appropriation actually happens, then an authentic personal identity emerges for someone from the basic metaphysical fact of her real personal identity, And then she thereby, at least partially or to some degree, achieves authenticity in an Existentialist sense, but also more specifically in my contemporary Kantian sense—that is, she thereby at least partially or to some degree achieves principled authenticity—and it is a truly sublime normative achievement. In addition to the 19th and 20th century Existentialists and some contemporary Kantians, also the classical Stoics, Augustine, Pascal, Spinoza, and some contemporary New Compatibilists in the tradition of Frankfurt's theory of free will and personhood[508] have all made similar points.

In turn, once these points have been philosophically recognized, they are easy enough to write down. But it must be stressed that *actually* achieving an authentic personal identity, and thereby at least partially or to some degree achieving principled authenticity via centripetal or centrifugal appropriation, is often immensely difficult to manage as a personal and practical matter. For there is poor tragic Sisyphus, painfully pushing that awful rock up the mountain, over and over again throughout all eternity, and yet as Camus puts it in his famous essay's stunning final paragraph:

> I leave Sisyphus at the foot of the mountain! One always finds one's burden again. But Sisyphus teaches the higher fidelity that negates the gods and raises rocks. He too concludes that all is well. This universe henceforth without a master seems to him neither sterile nor futile. Each atom of that stone, each mineral flake of that night-filled mountain, in itself forms a world. The struggle itself toward the heights is enough to fill a man's heart. One must imagine Sisyphus happy.[509]

Or, less cosmically, although perhaps equally poignantly, consider the problems faced by Williams's well-known fictional "lorry driver, who through no fault of his [own], runs over a child,"[510] and by the real-life Ketema Ross, who beat two old people with a broomstick during his schizophrenic episode. Sisyphus, the unfortunate truck driver, and Ketema Ross after his recovery must all *appropriate* some thoroughly nonideal facts about the world and about themselves, in order to achieve an authentic personal identity, at least partially or to some degree. These cases also show that someone's taking deep moral responsibility for *X* is perfectly consistent with her full recognition that *X* has happened *through no fault of her own*. Deep moral responsibility, therefore, can be *supererogatory* moral responsibility—freely taking upon oneself certain attitudes, choices, and acts that go creatively and substantially beyond what is merely morally required, in full recognition of the fact that, in that context, they are *not* morally required. Sisyphus does this by concluding that "all is well," and continuing to push his rock up that mountain forever. And perhaps the unfortunate truck driver does this by trying to help the family of the child he has accidently and blamelessly killed, and by subsequently creatively and substantially changing his life in some other ways—perhaps he gives a substantial part of his income to UNICEF, or becomes a part-time worker at a safe house/shelter for homeless mothers and children. Or perhaps, like the dying civil servant Kanji Watanabe in Akira

Kurosawa's iconically Existentialist 1952 film *Ikiru*, he devotes himself entirely to creating a children's playground on the site of a fetid swamp. And in the real world, Ketema Ross has been trying to change his life creatively and substantially since his release from confinement, by legally advocating on behalf of other people who have committed crimes "by reason of insanity."

At first, for most of us, it is cognitively difficult to imagine how these sorts of supererogatory acts of taking-deep-moral-responsibility could ever be really and truly performed by real-world people. But in fact, as Larissa MacFarquhar's *Strangers Drowning* vividly demonstrates, *some real-world people really and truly choose and do these things*. Moreover, I think that the cognitive failure of imagination that most of us experience here is in fact nothing but a cognitive illusion, induced by a wicked one-two punch combination of

> (i) the false presupposition *that all happiness is exclusively shallow happiness*, namely pleasure or preference-satisfaction via the satisfaction of first-order desires, together with
> (ii) the Enlightenment-induced Hobbesian cognitive myth of universal human egoism and mutual antagonism.

But when real-world altruists take supererogatory deep moral responsibility for awful things over which they had and have little or no control, then to that extent they have a good will, and they subjectively experience the deep happiness, or "negative satisfaction in one's existence, in which one is conscious of needing nothing," that Kant calls "self-fulfillment" or *Selbstzufriedenheit* (see section 3.4 above). Now some people *actually do* choose and act in this way. And if they actually do, then some real human persons *really can*. Moreover, every Kantian real human person *ought* to choose and act in this way, at least sometimes. Therefore you and I really can and ought to, too.

This deep happiness defies comprehension in terms of Humean instrumental reasons, and it requires a Kierkegaardian leap of non-instrumental, passionate Kantian stoical rational moral faith. Being able to say that "all is well" in this sense is categorically not the same as an ordinary "feeling good about things" or an ordinary "feeling good about yourself." On the contrary, it is a wholehearted, activist Kantian version of what the classical Stoics call *ataraxia*, what the Buddhists call *nirvana*, or what Nietzsche calls *amor fati*. It is a state of mind in which The Dear Self finally "shrinks to an extensionless point," and *the whole wide world*, especially including all the other real persons in The Realm of Ends, and including all the things that ought to be chosen and done with them and for them, appears in its place. In such a state of mind, you do not merely *passively accept* things exactly as they are: instead, you actively affirm or, in effect, *actively love* things exactly as they are, and thereby change your life around them—as per Wallace Stevens's man with the blue guitar:

> The man bent over his guitar,
> A shearsman of sorts. The day was green.
>
> They said, "You have a blue guitar,
> You do not play things as they are."
> The man replied, "Things as they are
> Are changed upon the blue guitar."

> And they said then, "But play, you must,
> A tune beyond us, yet ourselves,
>
> A tune upon the blue guitar
> Of things exactly as they are."[511]

In a seeming paradox, *things as they exactly are, are changed upon the blue guitar*. Or in Wittgenstein's terminology, the world is all that is the case, and all the facts can stay exactly the same, yet the world of the happy person is a different world from that of the unhappy person. In other words, this sublime emotional phenomenology is nothing more and nothing less than the subjective experience of an autonomous, higher-level, or Kantian real person's own principled authenticity in action, and indeed partially or to some degree attained.

***

> Psychological Spectrum:
>
> > [Callous Neuro-surgeon is a] single case in which, after a few changes, there will be psychological continuity. I shall discuss a *spectrum*, or range of cases, each of which is very similar to its neighbours. These cases involve all of the possible degrees of psychological connectedness.… In the case at the far end, the surgeon would cause very many switches to be simultaneously flipped. This would cause there to be no psychological connections between me and the resulting person. This person would be wholly like Napoleon. In the cases at the near end, the surgeon would cause to be flipped only a few switches. If he flipped the first switch, this would merely cause me to lose a few memories, and to have a few apparent memories that fit the life of Napoleon. If he flipped the first two switches, I would merely lose a few more memories, and have a few more of these new apparent memories. Only if he flipped all of the switches would I lose all my memories, and have a complete set of Napoleonic delusions.[512]

According to The Minded Animalist Criterion of Personal Identity, I clearly do survive whenever the "enough" conditions in the biophenomenological continuity condition (iiia) and the biological/neurobiological contintuity condition (iiib) are both satisfied. And clearly I do *not* survive whenever either of the "enough" conditions in The Minded Animalist Criterion (iiia) or particular members of some particular pair of cases in the psychological spectrum, I myself go out of existence forever and pseudo-Napoleon begins. In this connection, Parfit remarks that

> [i]t is hard to believe both that I would survive in one of these cases, and that, in the next case, I would cease to exist. Whether I continue to exist cannot be plausibly thought to depend on whether I would lose just a few more memories, and have a few more delusory memories, and have my character changed in some small way.[513]

On the contrary, however, as I reach and pass the precise threshold between satisfying The Minded Animalist Criterion of Personal Identity (iiia) and then not satisfying it, then I simply go out of existence forever.

We can look at it this way. Suppose that I am fainting. As I faint, I also necessarily pass through some series of definite degrees between consciousness and unconsciousness. So at some particular point between the particular members of some particular pair of cases in the biophenomenological spectrum of degrees of intensity of consciousness, I pass some precise threshold and suddenly become unconscious—just as my laptop computer suddenly goes into hibernation mode after a few minutes of not being used—and my occurrent consciousness thereby suddenly disappears. To be sure, the continuity of my consciousness is guaranteed by the fact that I remain dispositionally conscious, due to the continued existence of the neurobiological basis, or natural matrix, of my conscious states. But apart from that, how is this natural phenomenon of what Kant rather inelegantly called "elanguescence" (*CPR* B414) relevantly different from crossing a precise psychic threshold between being me and becoming pseudo-Napoleon, in the variant on Callous Neuro-surgeon in which both my pre-reflective/first-order consciousness and also my self-reflective/self-conscious consciousness are wiped clean and replaced with a pain-free Napoleonic biophenomenology?

The correct answer is that it is *not* relevantly different. Precisely somewhere or another along the spectrum of cases, the biophenomenological foundation of my life breaks up like a dropped cell-phone call, or a dropped Skype connection, my life thereby suddenly disappears forever, and pain-free pseudo-Napoleon's life begins. Presumably—assuming that I am fortunate enough and wholehearted enough to have a natural, timely death and not an accidental, untimely one[514]—then my own natural death will be just like that, only without any psychic successor, and without any biological-neurobiological continuant.

But as I noted above, even for my natural, non-accidental death there is still also the prospectively sad-seeming possibility that, like the unfortunate Iris Murdoch, my real personal life will literally end in my quasi-death. And in turn this would be followed by a post-personal psychic successor who shares some minimal degree of biophenomenological continuity with me, and is also a post-personal biological/neurobiological continuant who bears my proper name, but who himself is not actually me and will never actually be me, and instead is nothing but a has-been-me. It therefore seems undeniably true that at some particular point, Iris Murdoch literally quasi-died, such that a psychological continuant and her own body both lived on as jointly constituting her post-person. By the same token, Parfit's judgment about Psychological Spectrum is refuted, and Minded Animalism is further confirmed.

***

Brain Transplant:

> Suppose … that I am one of a pair of identical twins, and that both my body and my twin's brain have been fatally injured. Because of advances in neuro-surgery, it is not inevitable that these injuries will cause us both to die. We have between us one healthy brain and one healthy body. Surgeons can put these together…. Let us suppose … that the surgeons are able to connect my brain to the nerves in my twin's body. The resulting person would have no paralysis, and would be completely healthy. Who would his person be? This is not a difficult question…. If all my brain continues to exist and to be the brain of one living person, who is psychologically continuous with me, I continue to exist.[515]

According to The Minded Animalism Criterion of Personal Identity (iiib), namely, the biological-neurobiological continuity condition, I clearly do *not* survive the transplantation of my brain into my twin's body. This is because, according to The Essential Embodiment Theory and Minded Animalism, an essentially embodied mind is necessarily and *completely* biologically and neurobiologically embodied, that is, it must be embodied not only in the higher brain (cerebrum) and brain stem (cerebellum) of the living animal, but also in its nervous system, endocrine system, immune system, and cardiovascular system, right out to the skin. As complete, this biological-neurobiological embodiment can never be transplanted, since there is nothing left to transplant it *into*, although, to be sure, it could, thought-experimentally, be fully replicated, as in Simple Teletransportation. But replication in that teletransporting sense would, again, entail that I do *not* survive.

Notice, however, that The Minded Animalism Criterion of Personal Identity (iiib) does not imply that no changes can be made to someone's body without killing that person. As long as there *is enough* of the same complete biological-neurobiological embodiment, then my real personal identity can still be preserved under various changes. Here the phrase "enough of the same complete neurobiological embodiment" can be interpreted to mean "whatever constitutes the same total set of vital biological-neurobiological bodily systems," where *vital bodily systems* are just those systems whose permanent shut-down would normally cause the death of the entire organism. So, any bodily change that causally preserves the constitutive operations of all the vital bodily systems—up to and including various amputations, organ replacements, and certain kinds of permanent brain damage—also permits the continued survival of the person who is necessarily and completely neurobiologically embodied by the total set of such systems. Thus, like the tragically unfortunate protagonist of Dalton Trumbo's starkly powerful 1939 anti-war novel, *Johnny Got His Gun*, or like the somewhat less tragically unfortunate protagonist of Paul Voerhaven's weirdly compelling 1987 dystopian science fiction film, *Robocop*, the first in the *Robocop* series, at least in principle, I could really personally survive losing a limb, going blind, becoming deaf, losing my sense of smell, taste, or touch, the removal of one of my kidneys, a kidney transplant, a liver transplant, a heart transplant, commissurotomy, lobotomy, and various sorts of significant brain damage up to and including the destruction of one of my brain hemispheres, as long as it were replaced by something that causally preserved the integrity of the constitutive operations of my vital systems. Hence, at least in principle, I could also, as the same real human person, survive a hemispherectomy in that special causal preservationist sense.

This raises a crucial point. Gradual replacement of my body parts over time is allowed under The Minded Animalism Criterion of Personal Identity (iiib), provided that the continuity of all my vital bodily systems, *as* vital bodily systems, is ensured. Thus a real human person's necessary and complete biological/neurobiological embodiment is embodiment in a unified set of self-organizing complex thermodynamic living organismic systems that are severally constituted by their specific causal powers and goal-directed operations, and also by their specific roles in a real person's overall biological/neurobiological constitution. Necessary and complete biological/neurobiological embodiment is, therefore, not a merely *compositional* embodiment in a certain aggregate of parts made out of a certain kind of stuff. Hence classical Theseus's-ship type examples—whereby there is a gradual and eventually complete proper-part-by-proper-part replacement of something's compositional stuff, although the basic causal powers of each of those proper parts are preserved under each replacement—as applied to my living organismic body, are not only really possible, but also are *not* counterexamples to my

biological/neurobiological continuity under The Minded Animalism Criterion (iiib). In point of fact, I have already survived the gradual natural organismic modification and compositional replacement of all of my body parts, and the frequent replacement of my skin, as have you. But none of us could survive the permanent disruption or shut-down of our higher or lower brain, nervous system, endocrine system, immune system, or cardiovascular system, or indeed the permanent destruction (without an equivalent causal-functional replacement) of our skin—which we should therefore think of as an epidermal self-organizing, organismic thermodynamic *sub*-system, the outer membrane of our necessary and complete biological-neurobiological embodiment, and not as a mere static compositional envelope, aka "a bag of bones."

Thus Brain Transplant vividly brings out how The Minded Animalism Criterion of Personal Identity's requiring both biophenomenological and biological-neurobiological continuity sets it importantly apart from the classical approaches to personal identity. As I noted in section 6.1, these have almost universally fallen into one of two opposed categories: either the Psychological Approach, which says that some psychological relation is necessary and sufficient for diachronic personal identity, or the Somatic Approach, which says that on the contrary no psychological relation is necessary and sufficient for our persistence and that instead some fundamentally physical relation is necessary and sufficient. In this connection, Olson aptly notes the following:

> Here is a test case. Imagine that your cerebrum—the upper brain thought to be chiefly responsible for your mental features—is transplanted into my head. Two beings result: the person who ends up with your cerebrum and your mental features, and the empty-headed being left behind, which may still be alive but will have no mental features. If you would be the one who gets your cerebrum, that is presumably because some relation involving psychology suffices for you to persist, as the Psychological Approach says. If you would be the empty-headed vegetable, your identity consists in something non-psychological, as the Somatic Approach has it.[516]

Now according to The Minded Animalism Criterion, given this test case, clearly you are *neither* the one who gets the cerebrum *nor* the empty-headed vegetable. This because the empty-headed vegetable fails The Minded Animalism Criterion (iiia), namely, the biophenomenological condition, while the one who gets the cerebrum fails The Minded Animalism Criterion (iiib), namely, the biological-neurobiological continuity condition. In short, according to The Minded Animalism Criterion, clearly you die on *either* option of the test case, and it therefore follows that The Minded Animalism Criterion of Personal Identity is importantly distinct from each of the two classical approaches. This distinctness thereby holds even quite apart from The Minded Animalism Criterion's crucially important supplementary inclusion of The Minded Animalism Criterion (iiib), namely, the freedom-dominance condition, which itself, curiously enough, is completely overlooked by both of the classical approaches.

Or perhaps not so very curiously after all. This is because it does certainly seem as though the Psychological criterion as Parfit construes it, and also the Somatic criterion as Olson construes it, are both, in effect, intentionally designed so as to conform metaphysically to the doctrine of Compatibilism plus Soft Determinism. And this doctrine is clearly falsified by each of the four arguments for Local Incompatibilism with Respect to Natural Mechanism (see section 4.5 above and also Argument 4 in section 7.3).

***

My Division:

> I assume that I am one of three identical triplets…. My body is fatally injured, as are the brains of my two brothers. My brain is divided, and each half is successfully transplanted into the body of one of my brothers. Each of the resulting people believes that he is me, seems to remember living my life, has my character, and is every other way psychologically continuous with me. And he has a body that is very like mine.[517]

According to The Minded Animalism Criterion of Personal Identity condition (iiib), namely, the biological-neurobiological continuity condition, just as in the case of the whole brain transplant, again clearly I do *not* survive either of the hemispheric transplants. The consciousness of a real human person must be biologically-neurobiologically embodied in *all* of its vital bodily systems; the brain system is one of those vital bodily systems; therefore transplanting either my left brain hemisphere or my right brain hemisphere into some other body cannot suffice for my biological/neurobiological continuity.

To be sure, if we concede to Parfit that this division case is medically possible, then two different people exist after the accident and transplants—and I will call them, following Parfit's lead, "Lefty" and "Righty." But neither Lefty nor Righty is *me*, despite their having psychologies that replicate mine, and despite their having bodies that are very similar to mine. Lefty and Righty are new people. The accident and transplant brought about their creation. At best, they are *replicas* of me.

Think of it this way. Xerox copies or scanned .pdfs of some orginal, real-world hard copy document are not the same as the original document. Instead they are just newly created replicas of the original document. The original document is what it is, and not another thing. Often the difference between the original and the several replicas does not matter very much. But sometimes the original vs. replica difference does matter quite a lot. For example, not altogether unreasonably, the US government goes to considerable expense and trouble to preserve the original US Constitution. The same thing goes, with appropriate changes made for differences in context, for art galleries and most original works of art. So by analogy, Lefty and Righty are just replicas of my life. Now replicas of my life, like Lefty and Righty, are correspondingly *not* the owners of my original life, which is the sole possession of its author, namely, me.

It can be easily conceded by me that as an existential Kantian ethicist, and also as a real human person, a crooked timber, and "human, all too human," who would quite naturally have some selfish, egoistic, hedonistic, or consequentialist interest in replicas of his own biophenomenology and biology-neurobiology, I would no doubt have a non-trivial degree of special concern for Lefty and Righty and for their careers. Presumably I would wish them all the best, and also help them if I could by making provisions for them in my will, and so-on. Possibly, if, counterfactually, I were an exceptionally altrustic person, or perhaps, if, again counterfactually, I were a serious classical act-Utilitarian, then I might well be prepared to lay down my life for them right now and thereby bring Lefty and Righty into existence. And possibly, thinking now again of The Trolley Problem, it might also be morally permissible in some special circumstances for someone else to kill me in order to bring Lefty and Right into existence. But just because I naturally have some non-trivial degree of special concern for my

replicas, and just because I might freely (or be permissibly forced to) lay down my life for them, obviously it would not follow that either of them really *is* me.

From an existential Kantian ethical point of view, I can be deeply interested in and deeply care for other rational animals or real persons (whether via moral respect or merely via special concern) even though they are not in any substantive sense personally *identical* to me. It is true that, on my contemporary Kantian approach to the real metaphysics of real personhood, we all share the same basic innately specified online psychological and moral capacities. That is what guarantees our belonging to the same universal intersubjective community of real persons under constraints of mutual equal consideration—The Realm of Ends.[518] But even though we must all live together in this morally substantive sense, nevertheless we do not thereby share literally the same life. In order to do that, we would have to share literally the same living animal body. But given The One-Living-Body-Per-One-Real-Person-and-One-Real-Person-Per-One-Living-Body Thesis (see section 6.2), that is non-logically, essentially non-conceptually, synthetic a priori, or "strongly metaphysically" impossible.

***

Fifth, Parfit is wrong that what fundamentally needs to be explained about the fact of personal identity are

> (i) the unity of consciousness at a time, and
> (ii) the unity of a whole conscious life that also includes the unity of consciousness at any time as a necessary condition.

This is wrong in two different ways.

First, it is false that the unity of consciousness at a time, or the synchronic unity of consciousness, is a necessary condition of real personal identity. As Parfit correctly notes, neo-commissurotomy cases are logically, conceptually, analytically consistent, or "weakly metaphysically" (although Parfit does not use this particular label, which belongs to the philosophical jargon of the emerging Analytic metaphysics tradition in the 1980s and 1990s) consistent with the division or fission of consciousness into two branching psychological streams that are replicas of my conscious life, each of which depends on a single left or right brain hemisphere. But this does not entail that *I* would survive such a division. Indeed, according to The Minded Animalism Criterion of Personal Identity, such a division would clearly kill me. Moreover, Parfit neglects to note something important that I mentioned above, namely, that

> (i) actual neo-commissurotomy patients, apart from the visual field experiments mentioned, are otherwise behaviorally identical with normal-brained patients, and
> (ii) people with agenesis of the corpus callosum perform at the same level as normal-brained people on the visual field experiments.

In addition, the common experimental technique of "masking" in experimental psychology shows that concentration on one source of visual information, followed in rapid succession by another, tends to occlude conscious awareness of the second source, even though the second one explains definite "priming" results in later cognitive tasks. And it is a commonplace that

almost everyone, to some extent, is capable of divided attention with self-conscious focus on only one of several effectively performed tasks (aka "multi-tasking")—for example, skillfully driving on the interstate highway and also drinking hot coffee without spilling it, while at the same time thinking intensely about philosophy.

So the bottom line here is that the universality and necessity of the tripartite single-focus/multi-focal/non-focal structure of consciousness,[519] shows us that synchronic unity is *not* a necessary condition of consciousness. And by implication, in light of The Minded Animalism Criterion of Personal Identity's condition (iiia), namely, the biophenomenological continuity condition, the synchronic unity of consciousness is *not* a necessary condition of diachronic personal identity either. Subjective minimal coherence that is poised for free agency, and not unity, is the relevant necessary feature of consciousness.

Second, it is false that the unity of a real person's conscious life, whether synchronic *or* diachronic, is what fundamentally needs to be explained in an explanation of real personal identity. For example, some or perhaps even many people self-consciously experience their past lives as a series of definite phases or stages, each one of which is decisively put behind them as they move on to the next thing—for example, as Scarlett O'Hara (aka Vivien Leigh) famously remarks in the classic 1939 Hollywood epic melodrama directed by Victor Fleming, *Gone with the Wind*: "tomorrow is another day."

I am not saying that Scarlett O'Hara (or Vivien Leigh, for that matter) is (or was) a paragon of moral virtue, worthy of universal imitation. But Scarlett O'Hara does instantiate a genuine personality-type; and, frankly, there is also a certain amount of clear-sighted good moral-metaphysical common sense in what she says. For such people, it is not that they actually disown or forget their earlier phases or stages; it is rather just that they self-consciously experience themselves as significantly detached from, distanced from, and "beyond" those stages. You live, you make some choices and you do some things, and (as Nietzsche famously remarked) if they do not actually kill you, then you learn something from them that strengthens you, and afterwards you move on to the next thing. Indeed, an inability to move on to the next thing in this sense is the regrettable psychopathological phenomenon of nostalgia—as brilliantly explored, for example, in two entirely non-melodramatic Russian films directed by Andrei Tarkovsky in 1972 and 1983, *Solaris* and *Nostalghia*. But, in any case, the crucial point here is that even a fully Scarlett O'Hara-esque way of self-consciously experiencing one's own life over time does *not* undermine one's diachronic personal identity.

In one respect, the Scarlett O'Hara example is slightly misleading, since of course it is really possible that someone might really and truly have a diachronic unity of consciousness and at the same time also adopt an O'Hara-esque attitude towards her own past life. So the kind of case I am conceiving of is one in which a real person has an O'Hara-esque *self-conscious experience of her own past life, throughout that entire life* and not merely an O'Hara-esque *attitude, now, towards her own past life*. And that, I think, is really possible: more specifically, it seems to me really possible that a single real personal life possessing subjective minimal coherence poised for free agency can exist through a divided consciousness *over* time is just as real as the possibility that a single real personal life possessing subjective minimal coherence poised for free agency can exist through a divided consciousness *at* a time. All it would require, in effect, is getting up each and every morning, and wholeheartedly re-affirming to oneself: "*today* is another day."

Therefore, a *diachronic unity of consciousness* is not a necessary feature of a *diachronic real personal identity*, and not what fundamentally needs to be explained in an explanation of

real personal identity. What is relevantly necessary to diachronic real personal identity is not its *narrative* unity, but instead the pre-reflectively conscious awareness of one's being, in addition to whatever else one might be, an ongoing, subjectively minimally coherent, freedom-dominated biological-neurobiological and personal project that is, at least as far as our deepest aspirations are concerned, autonomously and wholeheartedly asymmetrically directed towards the future and towards one's own death, under the inherent guidance of personal projects and absolute moral principles. Or again, as Sartre puts it, the real person "exists only to the extent that he fulfills himself … [and] is nothing other than the ensemble of his acts, nothing other than his life."

So as against either the synchronic or the diachronic unity of consciousness, what fundamentally needs to be explained in the explanation of real personal identity is a synchronically and diachronically unique, intrinsically spatiotemporal, spatially oriented and thermodynamically irreversible, living organismic, freedom-dominated, aspirationally autonomous and wholehearted, conscious, intentional, caring subject in its necessary and complete biological-neurobiological embodiment. In view of my metaphysical analysis of the tripartite structure of free agency in chapter 5 above, the same point can be restated now even more succinctly. What fundamentally needs to be explained in the explanation of real personal identity is how *I am nothing more and nothing less than each and all of the stages of my whole essentially embodied, biological-neurobiological life-process from birth to death*, insofar as it is is immanently structured by my naturally growing and evolving innately specified capacities for veridical psychological freedom, deep freedom and deep (non-)moral responsibility, and principled authenticity.

Sixth and finally, Parfit is wrong that (real) personal identity is not what matters. In this connection, he says:

> [I]t may help to state, in advance, what I believe [My Division] to show. It provides a further argument against the view that we are separately existing entities. But the main conclusion to be drawn is that *personal identity is not what matters*.[520]

Minded Animalism is of course perfectly in line with Parfit's thesis that persons are not separately existing entities. According to The Minded Animalism Criterion of Personal Identity, real persons are essentially embodied minds, or minded animals, of a certain special kind—namely, the class of *2D rational* minded animals—and cannot be separated from their necessary and complete biological-neurobiological embodiment. But obviously it does not follow from this *either* that we are nothing but our bodies *or* that our real personal identity is not what matters.

On the contrary, real personal identity is *precisely* what matters, just as we always thought before we read Parfit, in view of the triadic or three-factor approach to real personal identity that is captured by The Minded Animalism Criterion of Personal Identity, along with the thesis that real personal identity is a special mereological relation of metonymous identity between proper spatiotemporal parts and unified spatiotemporal wholes. Insofar as one holds that real persons are essentially embodied minds and free agents, who are strictly identical with each and all the stages of their unique, finite, and complete lives, then at one and the same time one can also be a serious *non*-reductionist about real persons and about the absolute intrinsic value of their persistence, and also hold that real persons are *neither* separately existing entities *nor*

strongly supervenient entities—and thus also be a serious *non*-dualist and *non*-non-reductive physicalist about real persons.

Nevertheless, as I mentioned above, it is easily conceivable that I might have a special egoistic concern, or even moral respect, for people who are not identical to me, but in fact are replicas of me, as in Simple Teletransportation, Brain Transplant, and My Division. Indeed, it is also easily conceivable that I might have a special egoistic concern, or even moral respect, for a successor person who has merely most of the same living body as mine, like pseudo-Napoleon in Psychological Spectrum, or more intimately and poignantly, like my eponymous post-person in the final stages of Alzheimer's. That is, it is easily conceivable that I might come to feel about any of my replicas, or my post-person, as I already do about the people I love, my close friends, or my favorite students. In these cases, various psychological or physical similarities and other relations, whether genetic or socially produced, or even entirely accidental, are of considerable and even passionate importance to me, whether instrumentally or non-instrumentally, and thus the same sort of special egoistic concern or moral respect might well be extended to my replicas or to my post-person.

7.4 CONCLUSION

I conclude that we should reject Parfit's theory of personhood and personal identity, and accept Minded Animalism instead, along with The Minded Animalist Criterion of Personal Identity. Parfit's theory of reasons and persons is undoubtedly brilliant. But if Minded Animalism is correct, then Wollheim's *The Thread of Lif*e, published the very same year, 1984, although largely philosophically overshadowed by the success of Parfit's book, and largely neglected since then, was a more brilliant, deeper, and truer book than *Reasons and Persons*. Moreover, I also conclude that if The Minded Animalism theory of personhood and personal identity in chapters 6 and 7 is conjoined with the Natural Libertarianism theory of free agency in chapters 1 to 5, then we now have before us the basic elements of an intelligible, defensible, and unified real metaphysics of rational human free agency, real human personhood, and real human personal identity. More specifically, the fundamental connections between Minded Animalism and Natural Libertarianism are these: insofar as the basic conditions on our Kantian/Kierkegaardian deep freedom and our Kantian/Frankfurtian practical agency are satisfied, so too the basic conditions on our identity as real persons are satisfied, and conversely.

The central claim of this book, as I pointed out in section 1.3, is that the full metaphysical and normative power of a Kantian theory of human free will, practical agency, and persons, with deep and radical ethical and political implications downstream, can be captured by a correct understanding of how biological life in general, and the lives of human minded animals in particular, relate to the rest of physical nature. Or to capsulize the whole book, in ten words of two syllables or less: You *have* freedom-in-life, and you *are* your life.

REFERENCES

[1] S. Kierkegaard, "Either/Or, A Fragment of Life" in S. Kierkegaard, *The Essential Kierkegaard*, trans. H. Hong and E. Hong (Princeton, NJ: Princeton Univ. Press, 2000), pp. 37-83, at p. 72.

[2] L. Wittgenstein, *Philosophical Investigations*, trans. G.E.M. Anscombe (New York: Macmillan, 1953), 226e.

[3] D.R. Griffin, *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem* (Berkeley, CA: Univ. of California Press, 1998), p. 171.

[4] D. Hodgson, "Quantum Physics, Consciousness, and Free Will," in R. Kane (ed.), *The Oxford Handbook of Free Will* (Oxford: Oxford Univ. Press, 2002), pp. 85-110, at p. 86.

[5] H. Steward, *A Metaphysics for Freedom* (Oxford: Oxford Univ. Press, 2012), pp. 198-199.

[6] See, for example, S. Harris, *Free Will* (New York: Free Press, 2012).

[7] See J. Schuessler, "Philosophy That Stirs the Waters," *New York Times* (29 April 2013), available online at URL = <http://www.nytimes.com/2013/04/30/books/daniel-dennett-author-of-intuition-pumps-and-other-tools-for-thinking.html?emc=eta1&_r=0>.

[8] See, for example, the edgy 90s rock band, The Meat Puppets, "We Don't Exist," available online at URL = <https://search.yahoo.com/yhs/search?p=the+meat+puppets+we+don%27t+exist+youtube&ei=UTF-8&hspart=mozilla&hsimp=yhs-003>.

[9] See, for example, J. Ismael, *How Physics Makes Us Free* (Oxford: Oxford Univ. Press, 2016).

[10] Indeed, am I the only one to have noticed the stomach-turning unintentional similarity between the scientistic slogan "physics makes us free," and the hideously sanctimonious slogan posted over the gates of Auschwitz, Dachau, and other Nazi concentration camps, *Arbeit macht frei*?

[11] See, for example, Hanna, "Exiting the State and Debunking the State of Nature," THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.1; R. Hanna, "Radical Enlightenment: Existential Kantian Cosmopolitan Anarchism, With a Concluding Quasi-Federalist Postscript," in in D. Heidemann and K. Stoppenbrink (eds.), *Join, Or Die: Philosophical Foundations of Federalism* (Berlin: De Gruyter, 2016), pp. 63-90; and Hanna, *Kant, Agnosticism, and Anarchism*.

[12] The Minimal Law of Non-Contradiction says that *not every statement is both true and false*. See Hanna, *Cognition, Content, and the A Priori*, ch. 5.

[13] See, for example, R. Hanna, "Kant, the Copernican Devolution, and Real Metaphysics," in M. Altman (ed.), *Palgrave Kant Handbook* (London: Palgrave Macmillan, 2017), pp. 761-789.

[14] The leading figures of Analytic metaphysics include David Lewis, David Chalmers, Kit Fine, John Hawthorne, Ted Sider, and Timothy Williamson; and some of its canonical texts are Lewis's *On the Plurality of Worlds* (Oxford: Blackwell, 1986); Sider's *Writing the Book of the World* (Oxford: Oxford Univ. Press, 2011); Chalmers's *Constructing the World* (Oxford: Oxford Univ. Press, 2012); and Williamson's *Modal Logic as Metaphysics* (Oxford: Oxford Univ. Press, 2013).

[15] Sider, *Writing the Book of the World*, p. vii.

[16] See R. Hanna, *Kant and the Foundations of Analytic Philosophy* (Oxford: Oxford Univ. Press, 2001).

[17] See P.F. Strawson, *Individuals: An Essay in Descriptive Metaphysics* (London: Methuen, 1959); and P.F. Strawson, *Analysis and Metaphysics: An Introduction to Philosophy* (Oxford: Oxford Univ.Press, 1992).

[18] See, for example, F. Jackson, *From Metaphysics to Ethics: A Defense of Conceptual Analysis* (Oxford: Oxford Univ. Press, 1998).

[19] See, for example, W.V.O. Quine, "Epistemology Naturalized," in W.V.O. Quine, *Ontological Relativity and Other Essays*. New York: Columbia Univ. Press, 1969), pp. 69-90; W. Sellars, *Science, Perception, and Reality* (London: Routledge & Kegan Paul, 1963); and P. Maddy, *Second Philosophy: A Naturalistic Method* (Oxford: Oxford Univ. Press, 2007). And for a detailed critique of experimental philosophy/X-Phi, Hanna, *Cognition, Content, and the A Priori*, ch. 7.

[20] See R. Hanna and M. Maiese, *Embodied Minds in Action* (Oxford: Oxford Univ. Press, 2009), esp. chs. 1-2; and R. Hanna, "Minding the Body," *Philosophical Topics* 39 (2011): 15-40; then compare and contrast those with, for example, F. Jackson's highly influential "Epiphenomenal Qualia," *Philosophical Quarterly* 32 (1982): 127-136.

[21] See also P. Unger, *Empty Ideas: A Critique of Analytic Philosophy* (Oxford: Oxford Univ. Press, 2014).

[22] See Hanna, *Cognition, Content, and the A Prior*, esp. chs. 1-3.

[23] See also R. Hanna, "Life-Changing Metaphysics: Rational Anthropology and its Kantian Methodology," in G. D'Oro and S. Overgaard (eds.), *The Cambridge Companion to Philosophical Methodology*, (Cambridge: Cambridge Univ. Press, 2017), pp. 201-226.

[24] See also T. Horgan, "The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits," available online at URL = <http://www.humanamente.eu/PDF/Issue15_Paper_Horgan.pdf>. The crucial difference between the self-evident agentive phenomenolology I am talking about, and Horgan's agentive

phenomenology, *is that my agentive phenomenology is veridical*, and therefore it is what I call "locally incompatibilist," whereas Horgan's phenomenology is *non-veridical*, and indeed nothing but a classical-compatibilist/soft-determinist cognitive illusion. See also section 5.3 below.

[25] For detailed developments, defenses, and elaborations of the "mind-in-life" thesis, see E. Thompson, *Mind in Life* (Cambridge: Harvard Univ. Press, 2007); and Hanna and Maiese, *Embodied Minds in Action*.

[26] In chapter 2, I trace the origins of these non-reductive, immanent structuralist, dynamicist ideas back to Kant. But for some recent anticipations of basic aspects of this overall metaphysical conception, see I. Prigogine, *The End of Certainty* (New York; Free Press, 1997); Griffin, *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem*; and Timothy Dolch, "A Defense and Interpretation of the Causal Closure of the Physical," (PhD Dissertation, Univ. of Dallas, 2016).

[27] See also R. Wollheim, *The Thread of Life* (Cambridge. MA: Harvard Univ. Press, 1984).

[28] In *Embodied Minds in Action*, Maiese and I distinguish carefully between (1) "consciousness like ours" (or consciousness$_{lo}$), which is directly experienced by sentient living organisms like us, and (2) an unconstrained, unqualified notion of consciousness, which may include disembodied minds, angelic minds, divine minds, etc. In that book we focused almost exclusively on consciousness$_{lo}$ for various methodological reasons. In the present book I will focus my notion of consciousness in exactly the same way, but dispense with the slightly awkward subscripting convention.

[29] By *free volition*, or *minded animal agency*, I mean essentially the same thing that Helen Steward means by *animal agency*—see her *A Metaphysics for Freedom*, esp. chs. 1-2, 4, and 8. More generally, there are many significant parallels between Steward's metaphysics of freedom and mine, although it is free will and practical agency, that is, *rational* animal agency, that I am specifically focusing on here. But at the same time, it is a core feature of the metaphysics of freedom I am presenting, that the "locally incompatibilistic" fact of deep freedom, the "up-to-me-ness" or "ultimate sourcehood" of choices and acts, flows from the nature of essentially embodied, conscious, intentional, minded animal life itself. And that is the core of Steward's theory too, even despite its being rather unfortunately self-packaged and self-labeled by her as a new version of agent-causal Classical Libertarianism. In fact, her view is *much* closer to Natural Libertarianism than it is to agent-causal Classical Libertarianism.

[30] See Hanna, *Cognition, Content, and the A Priori*, esp. chs 2-3.

[31] In turn, the ethical theory developed in *Kantian Ethics and Human Existence* provides a crucial premise in the core argument for philosophical and political anarchism. See note 11 above.

[32] See note 11 above.

[33] See Hanna and Maiese, *Embodied Minds in Action*, esp. chs. 3-5.

[34] The neologisms "agential" and "agentive" both mean the same thing, namely, "directly concerned with, or characteristic of, intentional agents or intentional agency." Moreover, both terms are currently in use in the relevant philosophical literature: so I will use them interchangeably.

[35] See, for example, J. Campbell, M. O'Rourke, and D. Shier (eds)., *Freedom and Determinism* (Cambridge: MIT Press, 2004); J.M. Fischer, R. Kane, D. Pereboom, and M. Vargas, *Four Views on Free Will* (Oxford: Blackwell, 2007); R. Kane, *A Contemporary Introduction to Free Will* (Oxford: Oxford Univ. Press, 2005); Kane (ed.), *The Oxford Handbook of Free Will;* and G. Watson (ed)., *Free Will* (2nd edn., Oxford: Oxford Univ. Press, 2003).

[36] See, for example, C. Hoefer, "Causal Determinism," *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), E.N. Zalta (ed.), available online at URL = <http://plato.stanford.edu/archives/spr2016/entries/determinism-causal/>; M. McKenna and J. Coates, "Compatibilism," *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), E.N. Zalta (ed.), available online at URL = <http://plato.stanford.edu/archives/sum2015/entries/compatibilism/>.

[37] See, for example, R. Clarke and J. Capes, "Incompatibilist (Nondeterministic) Theories of Free Will," *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), E. N. Zalta (ed.), available online at URL = <http://plato.stanford.edu/archives/fall2015/entries/incompatibilism-theories/>; and K. Vihvelin, "Arguments for Incompatibilism," *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), E.N. Zalta (ed.), available online at URL = <http://plato.stanford.edu/archives/fall2015/entries/incompatibilism-arguments/>.

[38] This is a highly-compressed version of what are now called "manipulation arguments" for (source-) incompatibilism, following Derk Pereboom's fully-fleshed out classical "4-case" version in *Living Without Free Will* (Cambridge: Cambridge Univ. Press, 2001). But see also the post-classical, revised, and weakened version in D. Pereboom, *Free Will, Agency, and Meaning in Life* (Oxford: Oxford Univ. Press, 2014).

[39] See W. Sellars, "Philosophy and the Scientific Image of Man," in W. Sellars, *Science, Perception, and Reality* (New York: Humanities Press, 1963), pp. 1-40.

[40] For the distinction between formal mechanism and causal-nomological mechanism, see R. Hanna, "Kant's Anti-Mechanism and Kantian Anti-Mechanism," *Studies in History and Philosophy of Biological and Biomedical Science* 45 (2014), available online at URL = <http://www.sciencedirect.com/science/article/pii/S1369848614000107>. Gödel's incompletenesss theorems show that the thesis of universal formal mechanism does not work in mathematics or logic; see, for example, G. Boolos and R. Jeffrey, *Computability and Logic* (3rd edn.; Cambridge: Cambridge Univ. Press, 1989). Everyone accepts this. Nevertheless, the extension of Gödel's results to the thesis of universal causal-nomological mechanism in nature—which I will

call the thesis of "natural mechanism"—is controversial. See, for example, J.R. Lucas, "Minds, Machines, and Gödel," *Philosophy* 36 (1961): 112-127; J.R. Lucas, *The Freedom of the Will* (Oxford: Clarendon/Oxford Univ. Press, 1970), chs. 24-30; and R. Penrose, *The Emperor's New Mind* (Oxford: Oxford Univ. Press, 1990). I'm broadly sympathetic to the Lucas-Penrose project: but I don't think that Gödel's results *on their own* are sufficient to refute natural mechanism and/or explain free will. In fact, what's fully necessary and sufficient for the refutation of natural mechanism and/or the explanation of real (aka deep) freedom of the will are the four basic elements of the following philosophical package: (1) Gödel-incompleteness, PLUS (2) non-equilibrium complex systems thermodynamics, PLUS (3) the immanent structuralist metaphysics of the mind-body relation and intentional action that Maiese and I developed in *Embodied Minds in Action*, PLUS (4) the theory of essentially non-conceptual content and a priori knowledge that I developed in *Cognition, Content, and the A Priori*. See also chapter 2 below.

[41] See also Hanna and Maiese, *Embodied Minds In Action*, section 7.3.

[42] See also Hanna, *Kant, Science, and Human Nature*, section 8.3.

[43] See, for example, F. Varela, *Principles of Biological Autonomy* (New York: Elsevier/North-Holland, 1979); A. Weber and F. Varela, "Life After Kant: Natural Purposes and the Autopoietic Foundations of Biological Individuality," *Phenomenology and the Cognitive Sciences*, 1 (2002): 97-125; and Thompson, *Mind in Life*. Kant, Varela, and Thompson tend to identify *self-organizing* systems with *biological* systems. But while I fully agree that biological systems are indeed preeminent examples of self-organization, it is metaphysically important to recognize that self-organization in physical nature extends more widely than organismic activity, and can be found, for example, in quantum entanglement and other quantum effects, and in irreversible chemical processes like boiling water, not to mention macroscopically, in weather systems, traffic jams, etc., etc. In short, the metaphysical foundations of natural purposiveness and natural teleology can also be recognized in *physics* and *chemistry*, not only in biology.

[44] See, for example, Prigogine, *The End of Certainty*.

[45] See Hanna, *Kant, Science, and Human Nature*, chs. 3, 4, and 8; and also Hanna and Maiese, *Embodied Minds in Action*, chs. 6, 7, and 8.

[46] See, for example, E. Sosa and M. Tooley (eds.), *Causation* (Oxford: Oxford Univ. Press, 1993); and also J. Schaffer, D. Lewis, N. Hall, J. Collins, L. Paul, "Special Issue: Causation," *Journal of Philosophy* 97 (2000): 165-256.

[47] More precisely, CCP: Necessarily, all caused physical events have only event-causes that are consistent with all the deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or energy facts about the past, especially including The Big Bang.

So understood, CCP rules out any supernatural causes, and it also tells us what a thing's *physicality* is: having efficacious causal powers that are consistent with all the deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, together with all the settled quantity-of-matter-and/or energy facts about the past, especially including The Big Bang. CPP in this disambiguated and "precisified" sense fully allows for what Maiese and I in *Embodied Minds in Action*, chs. 6-7 call "jointly sufficient causes," i.e., essentially mental-and-physical causes that are complex singular event-causes exemplifying "dynamically emergent" biological properties and mental properties that involve "property-fusion" between fundamental mental properties and fundamental physical properties. The more precise contruals of CCP and dynamic emergence in the present book are principally due to the many important arguments and insights to be found in Dolch's "An Interpretation and Defense of the Causal Closure of the Physical."

[48] See, for example, Sellars, "Philosophy and the Scientific Image of Man," pp. 39-40.

[49] Strictly speaking, cosmological expansion and thermodynamic entropy aren't identical, although they do each constitute a temporal arrow pointing in the same direction as our subjective experience of time. See, for example, S. Hawking, *A Brief History of Time* (New York: Bantam, 1988), ch. 9. But for my purposes in this book, I'm treating cosmological expansion and entropy as parts of the same overall non-equilibrium, spatiotemporally asymmetric, unidirectional, complex thermodynamic process of the physical universe that began with The Big Bang.

It should also be noted that although The Big Bang belongs to the standard model of contemporary physics, there are also "finite but unbounded" models without real time and without singularities; see, for example, Hawking, *A Brief History of Time*, ch. 8. These, in turn, belong to a larger class of deterministic "block universe" models, without real time and its asymmetry; see, for example, Ismael, *How Physics Makes Us Free*.

[50] See, for example, D. Ratzsch and J. Koperski, "Teleological Arguments for God's Existence," *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), E.N. Zalta (ed.), available online at URL = <http://plato.stanford.edu/archives/spr2015/entries/teleological-arguments/>.

[51] This thesis is what, elsewhere, I call *radical agnosticism*. See, for example, R. Hanna, "If God's Existence is Unprovable, Then is Everything Permitted? Kant, Radical Agnosticism, and Morality," DIAMETROS 39 (2014): 26-69; and Hanna, *Kant, Agnosticism, and Anarchism*.

[52] See, for example, Hanna, "Kant, Nature, and Humanity," THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.2.

[53] Indeed, that's precisely why Marvin the Paranoid Android in Douglas Adams's classic pythonesque science fiction

saga, *Hitchhiker's Guide to the Galaxy*, is such an amusing character. See, for example, Wikipedia, "Marvin the Paranoid Android," available online at URL = <http://en.wikipedia.org/wiki/Marvin_the_Paranoid_Android>.

[54] For some recent attempts to reconstruct Kant's theory of freedom, see H. Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge Univ. Press, 1990); R. Hanna, *Kant, Science, and Human Nature* (Oxford: Clarendon/OUP, 2006), ch. 8; R. Hanna, "Kant, Causation, and Freedom," *Canadian Journal of Philosophy* 36 (2006): 281-306; R. Hanna, "Freedom, Teleology, and Rational Causation," in *Kant Yearbook* 1 (2009): 99-142; H. Hudson, *Kant's Compatibilism* (Ithaca, NY; Cornell Univ. Press, 1990); D. Pereboom, "Kant on Transcendental Freedom," *Philosophy and Phenomenological Research* 73 (2006): 537-567; E. Watkins, *Kant and the Metaphysics of Causality* (Cambridge: Cambridge Univ. Press, 2005); and A. Wood, "Kant's Compatibilism," in A. Wood (ed.), *Self and Nature in Kant's Philosophy* (New York: Cornell Univ. Press, 1984), pp. 73-101.

[55] Kierkegaard, "Either/Or, A Fragment of Life," p. 72.

[56] Kierkegaard, "Either/Or, A Fragment of Life," pp. 72 and 76.

[57] Kierkegaard, "Either/Or, A Fragment of Life," pp. 79-80. See also R. Clarke, "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism," *Mind* 118 (2009): 323-351, esp. 346-349.

[58] Forty years later, this remains absolutely true. I will add two other important qualifications, however. First, I distinguish sharply between (i) doing real philosophy, and (ii) being a professional academic philosopher who is paid by a private or public institution to teach, publish papers and books, do committee work and professional service, get along with obnoxious colleagues, mindlessly obey the rationally unjustified, coercive moralistic commands of administrators, etc., etc. It is not unreasonable to love (i) but despise (ii). Correspondingly, see also Hanna, "Thinking Inside and Outside the Fly-Bottle: The New Poverty of Philosophy and its Second Copernican Revolution," THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.4. Second, as the doomed Montgomery Clift character so correctly and wisely observes in *From Here to Eternity* (directed by F. Zinneman, 1953), just because you love something with all your heart, that doesn't mean it is ever going to love you back.

[59] There's also a very real sense in which my metaphysical conception of physical nature per se (hence also my approach to the philosophy of physics and chemistry) is throughly dynamicist: according to this conception, the natural world as a whole, including us, is nothing more and nothing less than a totality of non-equilibrium, complex thermodynamic systems with irreducible immanent structures of various kinds, that actualize potential energy, emerge in orientable actual space over irreversible actual time, realize the manifest world, and contribute to negentropy. This universal "no-levels," neutral monist, dual aspect, dynamicist metaphysical picture is collectively inspired by Aristotle's hylomorphic metaphysics (minus separable *noûs* or unmoved movers), by Kant's transcendental idealism and anti-mechanism (minus things-in-themselves), by Whitehead's process philosophy (minus atomism and theism), by existential phenomenology (plus an ethically-grounded conception of authenticity), and by Prigogine's non-deterministic interpretation of non-equilibrium thermodynamics (minus any confusion between non-mechanistic non-determinism and mechanistic indeterminism).

[60] D. Chalmers, *The Conscious Mind* (Oxford: Oxford Univ. Press, 1996), pp. 106-109, texts combined.

[61] M. Thompson, *Life and Action* (Cambridge, MA: Harvard Univ. Press, 2008), pp. 6 and 20, texts combined.

[62] Prigogine, *The End of Certainty*, p. 162.

[63] J. Searle, *Speech Acts* (Cambridge: Cambridge Univ. Press, 1969), p. 33.

[64] The representation of life presupposes the representation of physical motion, i.e., of *kinetic activity*, which I also think is an a priori formal representation, and, also just like the representation of life, a formal *essentially non-conceptual* representation. Thus the representation of motion, or kinetic activity, is in effect a "form of intuition" in Kant's sense. Moreover, just as we find it natural and scientifically useful to represent time, imagistically or diagramatically, in spatial terms, as a unidirectional line, or as a fourth dimension of a Euclidean manifold, so too we find it natural and scientifically useful to represent kinetic activity, imagistically or diagramatically, as what falls within the "light cone" of a four-dimensional Euclidean manifold. But all such images and diagrams, while mathematically-informative and factually correct, as far as they go, have a strong tendency to mislead by apparently implying that time and motion are *in themselves static facts*—hence "spatializing" them both, in Bergson's terminology. This in turn helps to motivate the deeply Parmenidean, deterministic, "block universe" conception of physical nature. On the contrary, however, *time and motion are no more inherently static than organismic life is*. They're all inherently *dynamic*. Thus time is primitively and essentially non-conceptually represented as a structured, continuous, content-neutral flow of "inner" conscious experiences; and motion is primitively and essentially non-conceptually represented as a structured, continuous, content-neutral flow of "outer" (i.e., proprioceptive, bodily) conscious experiences.

In this way, perhaps surprisingly, *meditation* and *free-style dancing* are philosophically more illuminating representations of time and motion than the mathematically-informative and factually correct diagrams one finds in contemporary introductory physics or cosmology texts. To be sure, this deeply important idea can also be found in Kant, the early Phenomenological tradition, and Bergson. But in any case, to keep things relatively simple in this book, I won't explicitly argue for the "transcendental aesthetic of kinetics" here.

[65] See, for example, J. Kim, *Supervenience and Mind* (Cambridge: Cambridge Univ. Press, 1993), esp. part 1; T. Horgan, "From Supervenience to Superdupervenience: Meeting the Demands of a Material World," *Mind* 102

(1993): 555-586; and Chalmers, *The Conscious Mind*, chs. 1-3.

[66] The recent and contemporary literature on constitution and grounding has been uniformly conducted under the (for me, clearly) false presupposition that noumenal metaphysics is defensible; see, for example, R. Wasserman, "Material Constitution," *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), E.N. Zalta (ed.), available online at URL = <http://plato.stanford.edu/archives/spr2015/entries/material-constitution/>; and R. Bliss and K. Trogdon, "Metaphysical Grounding," *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), E.N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2016/entries/grounding/>. According to my Kant-inflected view of real metaphysics and ontology, modally deep properties and relations all express *non-logical, essentially non-conceptual, or "strong metaphysical," synthetic a priori necessity*, which is essentially different from *logical, conceptual, or "weak metaphysical," analytic necessity*: hence my view is robustly *modal dualist*. See, for example, Hanna, *Kant and the Foundations of Analytic Philosophy*, chs. 3-5; Hanna and Maiese, *Embodied Minds in Action*, section 7.4; Hanna, *Cognition, Content, and the A Priori*, esp. chs. 2 and 4; and Hanna, "Kant, the Copernican Devolution, and Real Metaphysics." But to keep things relatively simple, I won't re-argue those (admittedly controversial) claims here.

[67] See, for example, A. Rosenberg and D. McShea, *Philosophy of Biology: A Contemporary Introduction* (New York: Routledge, 2008), esp. ch. 4; and A. Rosenberg and R. Arp (eds.), *Philosophy of Biology: An Anthology* (Chichester, UK: Wiley-Blackwell, 2009), esp. chs. 16-17.

[68] See, for example, M. Kaiser and B. Krickel, "The Metaphysics of Constitutive Mechanistic Phenomena," *British Journal for the Philosophy of Science* 67 (2016): 1-35, for a good survey of, and critical responses to, the standard literature.

[69] See, for example, G. Hunter, *Metalogic* (Berkeley, CA: Univ. of California Press, 1996), pp. 232-234.

[70] A. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society*, series 2, 42 (1936): 230-265, with corrections in 43 (1937): 644-546. See also, for example, G. Boolos and R. Jeffrey, *Computability and Logic* (3rd edn., Cambridge: Cambridge Univ. Press, 1989), ch. 3.

[71] See, for example, Boolos and Jeffrey, *Computability and Logic*, chs. 6-9, esp. pp. 52-54.

[72] A. Levy, "Machine-Likeness and Explanation by Decomposition," *Philosophers' Imprint* 14 (2014), also available online at URL = <http://www.arnonlevy.org/uploads/9/3/4/2/9342317/machines__decomp_phil_imprint_final.pdf>. The text cited is taken from the *PhilPapers* abstract, which is available online at URL = <http://philpapers.org/rec/LEVMAE>.

[73] See, for example, J. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (1980): 417-424; and J. Searle, *Minds, Brains, and Science* (Cambridge: Harvard Univ. Press, 1984).

[74] See, for example, Prigogine, *The End of Certainty*, pp. 66-67; and Wikipedia, "The Belousov-Zhabotinsky Reaction," available online at URL = <https://en.wikipedia.org/wiki/Belousov%E2%80%93Zhabotinsky_reaction>. The Belousov-Zhabotinsky reaction can be excited into self-organizing activity by means of the influence of light, using tris(bipyridine)ruthenium(II) chloride as a catalyst.

[75] See, for example, Boolos and Jeffrey, *Computability and Logic*, chs. 4-5.

[76] See, for example, Prigogine, *The End of Certainty*, pp. 163-182.

[77] See note 74 above.

[78] See, for example, Hanna, *Cognition, Content, and the A Priori*, esp. ch. 2.

[79] See R. Hanna, "On Kant's Term, 'Representation'," *Academia.edu*, available online at URL = <https://www.academia.edu/23421371/On_Kants_Term_Representation_>.

[80] Here is a point on which I deviate from Kant, somewhat: whereas for Kant, all natural purposes are organisms, for me, and other contemporary complex systems dynamicists, *not* all natural purposes are organisms—for example, the roiling movements of boiling water, traffic jams, and weather systems. See also, for example, J.A.S. Kelso, *Dynamic Patterns* (Cambridge, MA: MIT Press, 1995). Or in other words, some self-organizing systems are not organismic systems. Later in this chapter I will make a substantive proposal about what, over and above being self-organizing, being a living organism further requires.

[81] See, for example, Prigogine, *The End of Certainty*, ch. 6.

[82] See G.W.F. Leibniz, "The Principles of Philosophy, or, The Monadology," in R. Ariew and D. Garber (eds.), *Leibniz: Philosophical Essays* (Indianapolis, IN: Hackett, 1989), pp. 213-234, §17, p. 215.

[83] See J. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (1980): 417-424; J. Searle, "Intrinsic Intentionality: Reply to Criticisms of 'Minds, Brains, and Programs'," *Behavioral and Brain Sciences* 3 (1980): 450-456; and J. Searle, "The Chinese Room Revisited: Response to Further Commentaries on 'Minds, Brains, and Programs'," *Behavioral and Brain Sciences* 5 (1982): 345-348.

[84] See, for example, M. Hogarth, "Non-Turing Computers and Non-Turing Computability," in D. Hull, M. Forbes, and R. M. Burian (eds), *PSA 1994* (East Lansing, MI: Philosophy of Science Association, 1994), vol. 1, pp. 126–138; and M. Hogarth, "Deciding Arithmetic Using SAD Computers," *British Journal for the Philosophy of Science* 55 (2004): 681-691.

[85] See Wikipedia, "Malament-Hogarth Spacetime," available online at URL = <http://en.wikipedia.org/wiki/Malament-Hogarth_spacetime>.

[86] See, for example, G. Piccinini, "Computational Modeling vs. Computational Explanation: Is Everything a Turing Machine, and Does It Matter to the Philosophy of Mind?" *Australasian Journal of Philosophy*, 85 (2007): 93-115; G. Piccinini, "Computationalism, the Church-Turing Thesis, and the Church-Turing Fallacy," *Synthese*, 154 (2007): 97-120; G. Piccinini, "Computing Mechanisms," *Philosophy of Science*, 74 (2007): 501-526; G. Piccinini, "Computation Without Representation," *Philosophical Studies*, 137 (2008): 205-241;and G. Piccinini, "Computers," *Pacific Philosophical Quarterly*, 89 (2008): 32-73.

[87] See also Hanna, "Kant, Nature, and Humanity," THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.2.

[88] J. Mensch, *Kant's Organicism: Epigenesis and the Development of the Critical Philosophy* (Chicago, IL: Univ. of Chicago Press, 2013), pp. ix-x.

[89] Mensch, *Kant's Organicism*, p. 1.

[90] Mensch, *Kant's Organicism*, p. 27-28.

[91] Mensch, *Kant's Organicism*, p. 29.

[92] Mensch, *Kant's Organicism*, p. 50.

[93] Mensch, *Kant's Organicism*, p. 36.

[94] Mensch, *Kant's Organicism*, p. 61.

[95] Mensch, *Kant's Organicism*, p. 64.

[96] See I. Hacking, *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability and Statistical Inference* (Cambridge: Cambridge Univ. Press, 1975).

[97] T. Nagel, *Mind and Cosmos* (Oxford: Oxford Univ. Press, 2012), p. 17. The standard criticisms of Nagel (when they aren't simply ad hominem, or idealist-bashing) are (i) that he is ignorant of recent and contemporary work in evolutionary biology, and (ii) that, correspondingly, he overlooks the distinction between reductive and non-reductive biological (or more generally, scientific) naturalism. I think that these worries are philosophical red herrings, sophisms really, unintentionally or intentionally employed in order to avoid facing up to the deep organicist/anti-mechanist/liberal naturalist point that Nagel is trying to make. See also R. Hanna, "Nagel & Me: Beyond the Scientific Conception of the World," *Academia.edu* (2016), available online at URL = <https://www.academia.edu/4348336/Nagel_and_Me_Beyond_the_Scientific_Conception_of_the_World>.

[98] Nagel, *Mind and Cosmos*, p. 123.

[99] See R. Hanna, "Kant and Nonconceptual Content," *European Journal of Philosophy* 13 (2005): 247-290; R. Hanna, "Strong Kantian Non-Conceptualism," *Philosophical Studies* 137 (2008): 41-64; R. Hanna, "Beyond the Myth of the Myth: A Kantian Theory of Non-Conceptual Content," *International Journal of Philosophical Studies* 19 (2011): 321–396; R. Hanna, "Kant's Theory of Judgment," *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), E.N. Zalta (ed.), available online at URL = <https://plato.stanford.edu/archives/win2017/entries/kant-judgment/>, supplement 1; and Hanna, *Cognition, Content, and the A Priori*, ch. 2.

[100] See R. Hanna, "Kant's Non-Conceptualism, Rogue Objects, and the Gap in the B Deduction," *International Journal of Philosophical Studies* 19 (2011): 397-413; R. Hanna, "Kant, Hegel, and the Fate of Non-Conceptual Content," *Hegel Society of Great Britain Bulletin* 34 (2013): 1-32; and R. Hanna, "Kantian Madness: Blind Intuitions, Essentially Rogue Objects, and Categorial Anarchy," *Contemporary Studies in Kantian Philosophy* 1 (2016): 44-64.

[101] W. Sellars, "Empiricism and the Philosophy of Mind," in Sellars, *Science, Perception, and Reality*, pp. 127-196, at p. 173.

[102] See, for example, J.-J. Rousseau, *Reveries of the Solitary Walker*, trans. R. Goulbourne (Oxford: Oxford Univ. Press, 2011). The cited translations are by P. Harrison, available online at URL = <http://www.pantheism.net/paul/history/rousseau.htm>, seventh revery. Many thanks to Ericson Falabretti for reminding me about Rousseau's important influence on Kant's philosophy of nature, both human and non-human.

[103] W. Wordsworth, "My Heart Leaps Up," available online at URL = <http://www.poets.org/poetsorg/poem/my-heart-leaps>.

[104] P. Shelley, *Alastor*, available online at URL = <http://www.online-literature.com/shelley_percy/2778/>.

[105] M. Shelley, *Frankenstein; Or, the Modern Prometheus*, 1818 edn., available online at URL = <http://www.rc.umd.edu/editions/frankenstein>, vol. 1, ch. 3, underlining added.

[106] S. Alexander, "Natural Piety," in S. Alexander, *Philosophical and Literary Pieces* (London: Macmillan, 1939), pp. 299-315, at pp. 299, 310-311, and 306, underlining added.

[107] It is plausible to think that the analogies between the philosophy of biology and the philosophy of mind run very deep, principally because of the "mind-in-life" thesis which says that (i) conscious mind necessarily requires and includes biological life, and (ii) conscious mind and life are metaphysically continuous in the sense that the properties which are constitutive of biological life are also sufficient for consciousness, although in a more complex organizational structure, so that not every living thing is itself conscious. Whether the mind-in-life thesis is true or false, its meaningfulness suffices to show that the mind-body relation and the mind-life relation

are parallel structures, since no one denies that biological life is embodied. See, for example, Thompson, *Mind in Life*; and Hanna and Maiese, *Embodied Minds in Action*, esp. chs. 6-8.

[108] B. McLaughlin, "The Rise and Fall of British Emergentism," in A. Beckermann et al., (eds), *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism* (Berlin: De Gruyter, 1992).

[109] See, for example, Chalmers, *The Conscious Mind*.

[110] See for example, Kim, *Supervenience and Mind*.

[111] See, for example, Hanna, "Kant, Nature, and Humanity," THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.2.

[112] See also Hanna and Maiese, *Embodied Minds in Action*, section 8.2.

[113] As I mentioned in note 74 above, the Belousov-Zhabotinsky reaction *can* be excited into self-organizing activity via the influence of light, using tris(bipyridine)ruthenium(II) chloride as a catalyst. But even this still falls short of organismic life.

[114] For a theory of essentially non-conceptual content, and its epistemic, cognitive-semantic, and metaphilosophical implications, see Hanna, *Cognition, Content, and the A Priori*, esp. chs. 2, 4, and 8. See also Hanna, "Kant, the Copernican Devolution, and Real Metaphysics."

[115] T. Nagel, "What Is It Like To Be A Bat?," in T. Nagel, *Mortal Questions* (Cambridge: Cambridge Univ. Press, 1979), pp. 165-180. See also J. Levine, "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983): 354-361.

[116] D. Keleman, "British and American Children's Preferences for Teleo-Functional Explanations of the Natural World," *Cognition* 88 (2003): 201-221.

[117] T.S. Gendler, "Alief and Belief," *Journal of Philosophy* 105 (2008): 634-663; and T.S. Gendler, "Alief and Belief in Action (and Reaction)," *Mind and Language* 23 (2008): 552-585.

[118] Wittgenstein, *Philosophical Investigations*, §284, p. 98e.

[119] Wittgenstein, *Philosophical Investigations*, §357, p. 113e.

[120] Wittgenstein, *Philosophical Investigations*, p. 205e.

[121] Wittgenstein, *Philosophical Investigations*, p. 205e.

[122] Wittgenstein, *Philosophical Investigations*, p. 226e.

[123] Kelemen, "British and American Children's Preferences for Teleo-Functional Explanations of the Natural World," p. 216.

[124] Gendler, "Alief and Belief," pp. 637, 641, and 642.

[125] For an elaboration and defense of this conception of the a priori, see Hanna, *Cognition, Content, and the A Priori*, ch. 7.

[126] See, for example, T. Horgan and J. Tienson, "The Intentionality of Phenomenology and the Phenomenology of Intentionality," in D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings* (Oxford: Oxford Univ. Press, 2002), pp. 520-533.

[127] P. Godfrey-Smith, *Complexity and the Function of Mind in Nature* (Cambridge: Cambridge Univ. Press, 1996), p. 320.

[128] Thompson, *Mind in Life*, p. ix.

[129] As I indicated in sections 1.0 and 1.1, I'm also fully committed to the "life-in-energy" thesis, alongside the "mind-in-life" and "freedom-in-life" theses, according to this simple diagram of the basic metaphysical continuities:
free agency→conscious, intentional mind→organismic life→asymmetric matter/energy flows
But in this context, to keep things relatively simple, I'm not highlighting the life-in-energy thesis.

[130] See J. Maienschein, "Epigenesis and Preformationism," *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/epigenesis/>.

[131] See, for example, Hanna, "Kant's Anti-Mechanism and Kantian Anti-Mechanism."

[132] See also Hanna, "Kant, Nature, and Humanity," THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.2.

[133] See also Hanna, *Kant and the Foundations of Analytic Philosophy*, esp. chs. 4-5; and Hanna, *Kant, Science, and Human Nature*, ch. 6.

[134] See, for example, R. Hanna, "Kant and Nonconceptual Content," *European Journal of Philosophy* 13 (2005): 247-290; R. Hanna, "Kantian Non-Conceptualism," *Philosophical Studies* 137 (2008): 41-64; R. Hanna, "The Myth of the Given and the Grip of the Given," DIAMETROS 27 (March 2011), available online at URL = <http://www.diametros.iphils.uj.edu.pl/?l=2&p=anr25&m=25&if=0&ii=29&ik=27>; R. Hanna, "Beyond the Myth of the Myth: A Kantian Theory of Non-Conceptual Content," *International Journal of Philosophical Studies* 19 (2011): 321–396; and R. Hanna and M. Chadha, "Non-Conceptualism and the Problem of Perceptual Self-Knowledge," *European Journal of Philosophy* 19 (June 2011): 184-223.

[135] Hanna, *Cognition, Content, and the A Priori*, ch. 2.

[136] See, for example, J. Bermúdez and A. Cahen, "Nonconceptual Mental Content," *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), E.N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2015/entries/content-nonconceptual/>; G. Evans, *Varieties of Reference* (Oxford: Clarendon/Oxford Univ. Press, 1982), esp. chs. 4-6; and Y. Gunther (ed.), *Essays on Nonconceptual Content* (Cambridge: MIT Press, 2003).

[137] See, for example, J. McDowell, *Mind and World* (Cambridge: Harvard Univ. Press, 1994); McDowell, *Having the World in View*; S. Sedivy, "Must Conceptually Informed Perceptual Experience Involve Non-conceptual

Content?," *Canadian Journal of Philosophy* 26 (1996): 413-431; and B. Brewer, *Perception and Reason* (Oxford: Oxford Univ. Press, 1999).

[138] Maiese and I have worked out the relevant corresponding analysis of basic intentional actions, and also the relevant corresponding metaphysics of mental causation, in *Embodied Minds in Action*, chs. 3-5 and 6-8.

[139] See, for example, H. Dreyfus, "The Myth of the Pervasiveness of the Mental," in J. Schear (ed.), *Mind, Reason, and Being in the World* (London: Routledge, 2013), pp. 15-40, at p. 17; and S. Crowell, *Normativity and Phenomenology in Husserl and Heidegger* (Cambridge: Cambridge Univ. Press, 2013).

[140] See Hanna, *Cognition, Content, and the A Priori*, ch. 2.

[141] Actually, Kant's argument is unsound: there can be incongruent counterparts even if space and time are not transcendentally ideal in Kant's strong sense of transcendental ideality, but instead *manifestly real* in my sense. Nevertheless Kant's argument is still philosophically inspiring, because the existence of incongruent counterparts does indeed soundly entail that our discriminating representation of them is concept-independent: concepts do not suffice to discriminate the counterparts.

[142] See, for example, Godfrey-Smith, *Complexity and the Function of Mind in Nature*; E. Schrödinger, *What is Life?: The Physical Aspect of the Living Cell* (Cambridge: Cambridge Univ. Press, 1992); Varela, *Principles of Biological Autonomy*; B. Weber, "Emergence of Life and Biological Selection from the Perspective of Complex Systems Dynamics," in G. van de Vijver et al. (eds.), *Evolutionary Systems: Biological and Epistemological Perspectives on Selection and Self-Organization* (Dordrecht: Kluwer, 1998); and B. Weber and D. Depew, "Natural Selection and Self-Organization: Dynamical Models as Clues to a New Evolutionary Synthesis," *Biology and Philosophy* 11 (1996): 33-65.

[143] See, for example, S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution* (New York: Oxford Univ. Press., 1993); S. Kauffman, *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity* (New York: Oxford Univ. Press, 1995); G. Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium Systems* (New York: Wiley, 1977); and I. Prigogine, *Being and Becoming: Time and Complexity in the Physical Sciences* (New York: W.H. Freeman, 1980).

[144] B. Weber, "Life," *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), E.N. Zalta (ed.), available online at URL = <http://plato.stanford.edu/archives/spr2015/entries/life/>. See also Weber and Varela, "Life After Kant: Natural Purposes and the Autopoietic Foundations of Biological Individuality."

[145] See, for example, H. Caygill, *A Kant Dictionary* (Oxford: Blackwell, 1995), p. 214.

[146] See R. Hanna, "Mathematics for Humans: Kant's Philosophy of Arithmetic Revisited," *European Journal of Philosophy* 10 (2002): 328-353.

[147] See T. Skolem, "The foundations of elementary arithmetic established by means of the recursive mode of thought, without the use of apparent variables ranging over infinite domains," in J. Van Heijenhoort (ed.), *From Frege to Gödel* (Cambridge: Harvard Univ. Press), pp. 302-333.

[148] See, for example, Hunter, *Metalogic*, pp. 232-233.

[149] This thesis is what I call "weak or counterfactual transcendental idealism." See Hanna, *Cognition, Content, and the A Priori*, esp. section 7.3. Weak or counterfactual transcendental idealism is smoothly consistent with manifest realism, as per section 1.0 above.

[150] See, for example, K. Gödel, "On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems," in J. Van Heijenoort (ed.), *From Frege to Gödel* (Cambridge, MA: Harvard Univ. Press, 1967), pp. 596-617; and Boolos and Jeffrey, *Computability and Logic*, chs. 15-16.

[151] See, for example, M. Silberstein, and J. McGeever, "The Search For Ontological Emergence," *Philosophical Quarterly* 49 (1999): 182-200.

[152] See Prigogine, *The End of Certainty*, ch. 8.

[153] See Hanna and Maiese, *Embodied Minds in Action*, esp. chs. 3-5, and 8.

[154] On Kant's distinction between "regulative" and "constitutive" judgments, see Hanna, "Kant's Theory of Judgment," supplement 2.

[155] See, for example, J. Perry, "The Problem of the Essential Indexical," *Noûs* 13 (1979): 3-21.

[156] Chalmers, *The Conscious Mind*, p. 85.

[157] See, for example, E. Schrödinger, "The Present Situation in Quantum Mechanics: A Translation of Schrödinger's 'Cat Paradox' Paper, by J.D. Trimmer," *Proceedings of the American Philosophical Society* 124 (1980): 323-338.

[158] See Chalmers, *The Conscious Mind*, pp. 150-156.

[159] See J. Kim, "Mechanism, Purpose, and Explanatory Exclusion," in J. Kim, S*upervenience and Mind*, pp. 237-264; J. Kim, "The Myth of Nonreductive Materialism," in Kim, *Supervenience and Mind*, pp. 265-284; J. Kim, "The Non-Reductivist's Troubles with Mental Causation," in Kim, *Supervenience and Mind*, pp. 336-357; Kim, *Philosophy of Mind*, ch. 7; and J. Kim, *Physicalism, or Something Near Enough* (Princeton, NJ: Princeton Univ. Press, 2005), chs. 2-3.

[160] H. Jonas, *The Phenomenon of Life: Toward a Philosophical Biology* (Chicago, IL: Univ. of Chicago Press, 1966), pp. 75-76 and 79, texts combined.

[161] Correspondingly, I am doing the *real* metaphysics of the *manifestly real* world of *veridical appearances*, and neither the Analytic metaphysics of the noumenal world, nor analysis according to the Standard Picture, nor

scientific naturalism/experimental philosophy/second philosophy. See section 1.0 above.

[162] See, for example, C. Cleland and C. Chyba, "Defining 'Life'," *Origins of Life and Evolution of the Biosphere* 32 (2002): 387-393; C. Cleland and C. Chyba, "Does Life Have a Definition?, " in C. E. Cleland and M. A. Bedau (eds.), *The Nature of Life: Classical and Contemporary Perspectives from Philosophy and Science* (Cambridge: Cambridge Univ. Press, 2010), pp. 326-339; C. Cleland, "Life Without Definitions," *Synthese* 185 (2012): 125-144; and E. Machery, "Why I Stopped Worrying about the Definition of Life…And Why You Should as Well," *Synthese* 185 (2012): 145-164.

[163] See Hanna, *Cognition, Content, and the A Priori*, ch. 2.

[164] See, for example, E. Nagel, "Teleology Revisited: Goal-Directed Processes in Biology" and "Teleology Revisited: Functional Explanations in Biology," *Journal of Philosophy* 74 (1977): 261-279 and 280-301; and for an application and extension of the "systems-property" explanatory model to action theory, see H. Frankfurt, "The Problem of Action," in H. Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge Univ. Press, 1988), pp. 69-79.

[165] See J.O. de La Mettrie, *Man, A Machine* (Chicago, IL: Open Court, 1912), available online at URL = <https://archive.org/stream/manmachine00lame#page/n9/mode/2up>.

[166] See, for example, T. Huxley, "On the Hypothesis That Animals are Automata, and Its History," in D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings* (New York: Oxford Univ. Press, 2002), pp. 24-30.

[167] See Harris, *Free Will.*

[168] See, for example, Schuessler, "Philosophy That Stirs the Waters."

[169] See Thompson, *Mind in Life*; Hanna and Maiese, *Embodied Minds in Action*; and Nagel, *Mind and Cosmos*. In turn, roughly a century ago, four brilliant books all said basically the same thing: Alexander's *Space, Time, and Deity*; Bergson's *Creative Evolution*; Dewey's *Experience and Nature*; and Whitehead's *Process and Reality*.

[170] See, for example, H. Jonas, *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (Chicago, IL: Univ. of Chicago Press, 1984).

[171] See also Hanna, "Kant, Nature, and Humanity," THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.2.

[172] Ibid.

[173] Shelley, *Frankenstein; Or, the Modern Prometheus*, vol. 1, ch. 3.

[174] Prigogine, *The End of Certainty*, pp. 153-154.

[175] See T. Wilder, *Our Town: A Play in Three Acts* (New York: Harper & Row, 1938), available online at URL = <http://www.aasd.wednet.edu/cms/lib02/WA01001124/Centricity/Domain/74/Our_Town_full_text.pdf>.

[176] H. Frankfurt, "Freedom of the Will and the Concept of a Person," in H. Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge Univ. Press, 1988), pp. 11-25, at p. 19.

[177] D. Parfit, *On What Matters*, 2 vols. (Oxford: Oxford Univ. Press, 2011), vol. 1, p. xliv.

[178] See, for example, H. Frankfurt, "Identification and Wholeheartedness," in *The Importance of What We Care About*, pp. 159-176; and H. Frankfurt, *The Reasons of Love* (Princeton, NJ: Princeton Univ. Press, 2004).

[179] See, for example, R. Hanna, *Rationality and Logic* (Cambridge: MIT Press, 2006), esp. the Introduction, ch. 5, and ch. 7.

[180] See R. Hanna, "Rationality and the Ethics of Logic," *Journal of Philosophy* 103 (2006): 67-100.

[181] Voltaire, "La Bégueule," (1772), lines 1-2: *Dans ses écrits, un sage Italien Dit que le mieux est l'ennemi du bien.*

[182] See, for example, C. Korsgaard, *Self-Constitution: Agency, Identity, and Integrity* (Oxford: Oxford Univ. Press, 2009), pp. 31, 81, 91, and 160.

[183] F. Dostoevsky, *The Brothers Karamazov*, trans. D. Magarshack (2 vols., Harmondsworth, Middlesex: Penguin, 1958), vol. 2, p. 743.

[184] See, for example, J. Harris, *Of Liberty and Necessity: The Free Will Debate in Eighteenth Century British Philosophy* (Oxford: Clarendon/OUP, 2005).

[185] See note 54 above.

[186] See, for example, Allison, *Kant's Theory of Freedom*, pp. 47-53; Pereboom, "Kant on Transcendental Freedom"; Watkins, *Kant and the Metaphysics of Causality*, chs. 5-6; and Wood, "Kant's Compatibilism."

[187] See, for example, *Kant's Theory of Freedom*, ch. 13; and A. Wood, *Kant's Ethical Thought* (Cambridge: Cambridge Univ. Press, 1999), pp. 180-182.

[188] See note 159 above.

[189] Strictly speaking, The Biological Theory of Transcendental Freedom is just one part of a more comprehensive interpretation of Kant's theory of freedom that I call *The Embodied Agency Theory*. See Hanna, *Kant, Science, and Human Nature*, ch. 8.

[190] See R. Hanna, "Kant, Causation, and Freedom," *Canadian Journal of Philosophy* 36 (2006): 281-306. See also B. Hall, *The Post-Critical Kant* (London: Routledge, 2015).

[191] See sections 1.1 and 2.3 above. See also D.R. Griffin, *Untying the World-Knot*; G. Rosenberg, *A Place for Consciousness* (New York: Oxford Univ. Press, 2004), pp. 8-10.

[192] See Hanna, *Kant and the Foundations of Analytic Philosophy*, sections 2.3 to 2.4; Hanna, *Kant, Science, and Human Nature*, section 6.1; and Hanna, *Cognition, Content, and the A Priori*, section 7.3.

[193] Nagel, *Mind and Cosmos*, p. 17.

[194] Oddly enough, in the text I've elided, Nagel has "Plato and perhaps also…." But maybe he's thinking of the *Neo*platonists, who would indeed count as predecessors of the absolute idealists; or maybe he's thinking of McDowell's neo-Hegelian "naturalized platonism."

[195] See Hanna, *Cognition, Content, and the A Priori*, p. 341.

[196] See R. Hanna and E. Thompson, "Neurophenomenology and the Spontaneity of Consciousness," in E. Thompson (ed.), *The Problem of Consciousness* (Calgary, AL: University of Alberta Press, 2005), pp. 133-162.

[197] See, for example, R. Chisholm, "Human Freedom and the Self," in Watson (ed.), *Free Will*, pp. 26-37; R. Clarke, "Agent Causation and Event Causation in the Production of Free Action," *Philosophical Topics* 24 (1996): 19-48; and T. O'Connor, *Persons and Causes* (New York: Oxford Univ. Press, 2000).

[198] See, for example, Watkins, *Kant and the Metaphysics of Causality*.

[199] See section 2.3 above.

[200] See, for example, A. Breitenbach, "Two Views on Nature: A Solution to Kant's Antinomy of Mechanism and Teleology," *British Journal for the History of Philosophy* 16 (2008): 351-369; A. Breitenbach, "Teleology in Biology: A Kantian Perspective," *Kant Yearbook* 1 (2009): 31-56; H. Ginsborg, "Kant on Understanding Organisms as Natural Purposes," in E. Watkins (ed.), *Kant and the Sciences* (New York: Oxford Univ. Press, 2001), pp. 231-258; P. Guyer, *Kant's System of Nature and Freedom* (Oxford: Oxford Univ. Press, 2005), chs. 5 and 13; and J. Kreines, "The Inexplicability of Kant's *Naturzweck*: Kant on Teleology, Explanation, and Biology," *Archiv für Geschichte der Philosophie* 87 (2005): 270-311.

[201] Korsgaard, *Self-Constitution*, p. 39, underlining added.

[202] See G. Santayana, *Skepticism and Animal Faith* (New York: Dover, 1955).

[203] For Kant, laws do not have to be semantically insensitive to contextual conditions or mentalistic facts in order to be necessary and strict, since they can also be *non-logically or synthetically necessary, that is, restrictedly necessary*. See Hanna, *Kant and the Foundations of Analytic Philosophy*, ch. 5. Fodor calls such psychological laws "ceteris paribus laws": see his "Making Mind Matter More," in J. Fodor, *A Theory of Content and Other Essays* (Cambridge: MIT Press, 1990), 137-159. Where Kant and Fodor would strongly disagree is that for Kant, these synthetically necessary psychological laws are wholly particular and *one-time-only* or "one-off," not general, whereas for Fodor they must be general laws.

[204] See, for example, R. Hanna, "Mathematics for Humans: Kant's Philosophy of Arithmetic Revisited," *European Journal of Philosophy* 10 (2002): 328-353; and Hanna, *Kant, Science, and Human Nature*, ch. 6.

[205] See also K. Westphal, *Kant's Transcendental Proof of Realism* (Cambridge: Cambridge Univ. Press, 2004), pp. 229-243.

[206] See also Lucas, *The Freedom of the Will*; chs. 24-30; and Lucas, "Minds, Machines, and Gödel."

[207] Of course, I could be wrong that the Goldbach conjecture and the Continuum Hypothesis are logically unprovable: at the moment, they are merely unproven. But it *might* be that they are unproven *because* they are logically unprovable, and that this, in turn, is because pure spatiotemporal intuition is required for their meaning, their truth, and our knowing them a priori. In that case, they would be synthetic a priori truths, not analytic or logical truths, and their logical unprovability would be philosophically explained. For some tentative thoughts along these lines about the rational knowability and modal status of the Continuum Hypothesis, see Hanna, *Cognition, Content, and the A Priori*, section 8.2.

[208] —Together with my co-author, Michelle Maiese, that is: see Hanna and Maiese, *Embodied Minds in Action*, chs. 3-5. In that book, we develop a Frankfurt-inspired guidance-control theory of action that explicitly rejects the causal theory of action, as per Frankfurt's "The Problem of Action." But unlike standard Frankfurt-inspired theories of action, which are compatibilist/soft determinist—see, for example, J.M. Fischer and M. Ravizza, *Responsibility and Control* (Cambridge: Cambridge Univ. Press, 1998)—our theory is neo-Aristotelian, and based on non-supervenient, essentially embodied, pre-reflectively conscious structuring causes.

[209] See, for example, C. Korsgaard, *Creating the Kingdom of Ends* (Cambridge: Cambridge Univ. Press, 1996); C. Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge Univ. Press, 1996); C. Korsgaard, *Self-Constitution: Agency, Identity, and Integrity* (Oxford: Oxford Univ. Press, 2009); and H. Sidgwick, *The Methods of Ethics* (London: Macmillan, 1907), pp. 511-516.

[210] In fact there are several non-trivially different versions of Kantian Conceptualism. See R. Hanna, "Kant's Theory of Judgment," supplement 1. For example, *strong* versions of Kantian Conceptualism hold that the understanding not only necessarily but also sufficiently determines all intentional contents. But those differences don't matter for the point I'm making here.

[211] See Hanna, "Kant's Theory of Judgment."

[212] See, for example, I. Geiger, "Rational Feelings and Moral Agency," *Kantian Review* 16 (2011): 283-308; D. Guevara, *Kant's Theory of Moral Motivation* (Boulder, CO: Westview Press, 2000); R. McCarty, "Kantian Moral Motivation and the Feeling of Respect," *Journal of the History of Philosophy* 31 (1993): 421-435; R. McCarty, "Motivation and Moral Choice in Kant's Theory of Rational Agency," *Kant-Studien* 85 (1994): 15-31; J. Wuerth, *Kant on Mind, Action, and Ethics* (Oxford: Oxford Univ. Press, 2014); and J. Wuerth, "Sense and Sensibility in Kant's Practical Agent: Against the Intellectualism of Korsgaard and Sidgwick," *European Journal of Philosophy* 21 (2013): 1-36.

[213] See, for example, O. Ware, "Kant on Moral Sensibility and Moral Motivation," *Journal of the History of*

Philosophy 52 (2014): 727-746.

[214] For a different take on *Eigenliebe* and *Eigendünkel*, see O. Ware, "Self-Love and Self-Conceit in Kant's Moral Psychology," (Unpublished MS, 2013 version).

[215] See D. Hume, *Treatise of Human Nature*, 2nd edn., Oxford: Clarendon/Oxford Univ. Press, 1978), book II, part III, section iii, p. 416.

[216] J. Milton, "Paradise Lost," in J. Milton, *The Poems of John Milton* (2nd edn; New York: Ronald Press, 1953), pp. 204-487, act iv, lines 108-110.

[217] See, for example, Augustine, *Confessions*, trans. R.S. Pine-Coffin (Harmondsworth, Middlesex UK: Penguin, 1961), book VIII, ch. 5, pp. 164-165.

[218] See H. Arendt, *Eichmann in Jerusalem: A Report on the Banality of Evil* (revised & enlarged edn.; Harmondsworth, Middlesex: Penguin, 1977). Historical evidence uncovered since then indicates that the actual Eichmann was, in fact, near-Satanically evil himself, so I'm using *Arendt's Eichmann* as my example, not the actual Eichmann.

[219] Geiger, "Rational Feelings and Moral Agency," p. 303.

[220] See, for example, M. Slote, *From Enlightenment to Receptivity: Rethinking our Values* (Oxford: Oxford Univ. Press, 2013). Of course, I don't deny that the classical caricature of Kant as "a cold, dry, rationalist" makes a very convenient philosophical punching-bag and strawman. But it's not Kant.

[221] See, for example, B. Williams, "Internal and External Reasons," in B. Williams, *Moral Luck* (Cambridge: Cambridge Univ. Press, 1981), pp. 101-113.

[222] See, for example, H. Frankfurt, "Freedom of the Will and the Concept of a Person," in H. Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge Univ. Press, 1988), pp. 11-25; H. Frankfurt, "Identification and Wholeheartedness," in Frankfurt, *The Importance of What We Care About*, pp. 159-176; and Frankfurt, *The Reasons of Love*.

[223] See, for example, Allison, *Kant's Theory of Freedom*. The Incorporation Thesis is widely accepted in the contemporary Kant-literature. But it is metaphysically mysterious how a rational capacity that is essentially external to human desires can nevertheless act directly on desires and convert them into maxims. In effect, it is a local miracle of Agent-Causation that presupposes the philosophically implausible Timeless Agency Theory of freedom.

[224] This thesis, when fully unpacked, is the same as the action-theory that Maiese and I develop in *Embodied Minds in Action*, chs. 3-5: see also note 208 above.

[225] In this text, and in other related texts I quote from *Religion within the Boundaries of Mere Reason*, I have elided references to God. This is not because I think that the notion of God is unimportant in Kant's conception of practical rationality and moral agency—on the contrary, it is of fundamental importance—but only because Kant's moral theology is exceptionally subtle, even by Kantian standards. Hence discussing it here would only add needless complexity and length to the present account. In any case, I discuss Kant's moral theology in detail in R. Hanna, "If God's Existence is Unprovable, Then is Everything Permitted? Kant, Radical Agnosticism, and Morality," DIAMETROS 39 (2014): 26-69. For an alternative account of the same material, see A. Chignell, "Rational Hope, Moral Order, and the Revolution of the Will," in E. Watkins (ed.), *Divine Order, Human Order, and the Order of Nature* (Oxford: Oxford Univ. Press, 2013), pp. 197-218; and A. Chignell, *What May I Hope?* (London: Routledge, 2015).

[226] In such a case, moreover, it's entirely likely *that the police are acting in a rationally unjustified and immoral way too*. See Hanna, *Kant, Agnosticism, and Anarchism*, sections 3.8 and 3.10.

[227] Korsgaard, *Self-Constitution*, p. 160.

[228] See, for example, Augustine, "Enchiridion on Faith, Hope, and Love," in S. Cahn and P. Markie (eds.), *Ethics: History, Theory, and Contemporary Issues* (3rd edn., New York: Oxford Univ. Press, 2006), pp. 195-202.

[229] See, for example, Augustine, *Confessions*, book VIII, ch. 5, pp. 164-165.

[230] See, for example, J. Raz, "The Guise of the Bad," available online at URL = <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2595986>.

[231] See, for example, Raz, "The Guise of the Bad."

[232] He couldn't be *explicitly* committed to it, because his use of the term "moral evil" is broader than mine, and essentially equivalent to what I'm calling "moral disvalue or wrong."

[233] See T. Schapiro, "The Nature of Inclination," *Ethics* 119 (2009): 229-256.

[234] In *Kant, Agnosticism, and Anarchism*, parts 2 and 3, I focus on what I call the *real* realm of ends, which is the ethical community consisting of all actual, Earth-bound rational human animals in the process of striving towards the regulative Idea of a post-State, social anarchist condition, aka "the sole and complete good on Earth," aka "God's kingdom on Earth."

Confusingly, in the *Critique of Practical Reason* (*CPrR* 5: 110-111), Kant uses the term "the highest good" to mean the same as what he had called "the sole and complete good" in the *Groundwork*, namely the ethical totality consisting of human happiness proportioned to moral virtue (*GMM* 4: 396). The notion of this ethical totality, in turn, is the same as the notion captured by the third Postulate of Pure Practical Reason, the moral belief in God's existence as a regulative Idea, since only an all-powerful and all-good God could create such an ethical totality (*CPrR* 5: 124-132). By contrast, in the *Groundwork*, "the highest good" means *a good will* (*GMM* 4: 393-394), which is the same as what he calls "the supreme good" in the second *Critique* (*CPrR* 5: 111).

To be sure, all of these slightly different notions are essentially connected, since the supreme good (namely, a good will) is the formal essence and immanent structure of the sole and complete good (that is, the ethical totality consisting of human happiness proportioned to moral virtue) that only an all-powerful and all-good God could create. Finally, when this ethical totality is not merely a moral belief or regulative Idea of pure practical reason, but is actually implemented in the world, spread out over all of humanity, over the long haul, then it is *the sole and complete good on Earth*, or *God's kingdom on Earth*—or what, in *Kant, Agnosticism, and Anarchism*, I call *The Kosmopolis* and *Utopia Now*.

[235] In "The Nature of Inclination," Schapiro aptly calls this a "middle road" approach to the nature of the will, i.e., one that neither identifies the will with desire nor over-intellectualizes the will.

[236] Many thanks to Spencer Case for formulating this worry.

[237] This line of thinking is closely related to G.E. Moore's arguments in *Principia Ethica* against The Naturalistic Fallacy, and in support of the primitiveness and irreducibility of The Good. Correspondingly, in *Kantian Ethics and Human Existence*, section 1.4, I develop a new-and-improved version of Moore's notorious Open Question Argument that is specially reconfigured to fit my version of contemporary Kantian ethics.

Aside from the naturalistic fallacy, the *other* fundamental fallacy about rational normativity is the substitution of the 1D conception of rational normativity for the 2D conception. This fallacy, in turn, has two somewhat distinct meta-ethically untoward or even tragic sub-types: (i) thinking that any sort of moral disvlaue or wrongness, hence any sort of choice or action falling short of ideal goodness or rightness, entails the death of rational normativity, therefore non-agency, and therefore non-responsibility (Smerdyakov's fallacy), and (ii) thinking that mere *morally bad* choice or action is the same as *morally evil* choice or action (moral puritanism). Morally puritanical fallacious moral thinking clearly underwrites such moral abominations as, for example, Prohibition, drug-possession-and-use laws, "three strike" laws, deporting legal or illegal immigrant non-citizens for minor crimes, etc., etc.

[238] See Hume, *Treatise of Human Nature*, book II, part III, section iii, p. 416.

[239] See S. Budick, *Kant and Milton* (Cambridge, MA: Harvard Univ. Press, 2010).

[240] In terms of the theory of knowledge I developed *in Cognition, Content, and the A Priori*, "categorical epistemology," this Moorean paradigm of skepticism-proof knowledge counts as "High-Bar a posteriori knowledge." See Hanna, *Cognition, Content, and the A Priori*, esp. section 1.3 and ch. 3. Unlike Moore, however, I also think that all skepticism-proof, High-Bar knowledge presupposes a (suitably weak, counterfactual) version of transcendental idealism; see *Cognition, Content, and the A Priori*, chs. 6-8.

[241] B. Pascal, *Pensées*, trans. A.J. Krailsheimer (Harmondsworth, Middlesex, UK: Penguin, 1955), section 4, # 277, p. 154, translation modified slightly.

[242] See, for example, J. Sarasohn, *Science on Trial: The Whistleblower, the Accused, and the Nobel Laureate* (New York: St. Martin's, 1993). This is an account of the notorious "David Baltimore case."

[243] See Hanna, *Kant, Science, and Human Nature*, ch. 7.

[244] See R.M. Rilke, "Archaic Torso of Apollo," trans. S. Mitchell, in R.M. Rilke, *Selected Poetry of Rainer Maria Rilke* (New York: Vintage Books, 1989), pp. 60-61.

[245] *Selbstzufriedenheit* could also be translated as "self-gratification" or "self-satisfaction." But in English these have the unfortunate and misleading connotations of either masturbation (indeed this connotation in German was, not altogether suprisingly, made fun of by Nietzsche) or sanctimonious smugness. "Self-fulfillment" avoids these, and also captures the inherently teleogical character of Kant's idea, because it includes the notions of self-determination and self-realization.

[246] Kierkegaard, "Purity of Heart is to Will One Thing," in Kierkegaard, *The Essential Kierkegaard*, p. 271.

[247] Frankfurt, "Freedom of the Will and the Concept of a Person," p. 21.

[248] See, for example, Wikipedia, "The Daleks," available online at URL = <http://en.wikipedia.org/wiki/The_Daleks>. The analogy is not perfect, because Daleks are always malevolent, whereas inauthenticity can also be benign. But the analogy is *almost* perfect, since Daleks are living conscious creatures who have literally turned themselves into machines.

[249] Cf. Hume, *Treatise of Human Nature*, book II, part III, section iii, p. 415.

[250] This is the upshot of the third postulate, or "God postulate," in the Postulates of Pure Practical Reason section in the *Critique of Practical Reason*, and also the upshot of *Religion Within the Boundaries of Mere Reason*—which really should have been entitled *Religion Only Within the Limits of Pure Practical Reason*. See note 234 above; and also Hanna, *Kant, Agnosticism, and Anarchism*, esp. part 1.

[251] See Hanna, *Kant, Agnosticism, and Anarchism*, especially part 3.

[252] A. Schopenhauer, *The World as Will and Representation*, trans. E.F.J. Payne (2 vols., New York: Dover, 1969), vol. 1, §23, pp. 113-114.

[253] In *Cognition, Content, and the A Priori*, ch. 7, I distinguish *rational* intuitions from other things many contemporary philosophers call "intuitions" (for example, spontaneous mere opinions, aka "armchair judgments," and "intellectual seemings"), by defining them in a Kantian neo-rationalist way as analytically fallible, non-inferential, active takings of certain statements to be necessary and a priori. Then I divide rational intuitions into three classes: (i) authoritative, intrinsically compelling, or self-evident (essentially reliable), (ii) constructed or derived (fairly reliable), and (iii) prima facie (fairly unreliable). Then I distinguish between (i)

basic or fundamental authoritative intuitions, and (ii) non-basic or non-fundamental authoritative intuitions that presuppose the basic ones. I regard the intuitive definition of free will I present here as being, at the very least, the expression of a *constructed or derived* rational intuition; but I do also think that a case could be made for its being the expression of a *non-basic authoritative rational intuition*, built up as a conjunction of basic authoritative rational intuitions.

[254] See also S. Wolf, "Responsibility, Moral and Otherwise," *Inquiry* 58 (2015): 127-142. Wolf's own account of the nature of deep responsibility, however, like virtually all contemporary accounts of responsibility (see note 255 directly below), is *attribution-theoretic*. Therefore ultimately, what is for her the non-moral fact of deep responsibility would count as "shallow responsibility" in my sense of that term.

[255] See, for example, R. Clarke, M. McKenna, and A. Smith (eds.), *The Nature of Moral Responsibility* (Oxford: Oxford Univ. Press, 2015).

[256] See, for example, C. Siebert, "Should a Chimp Be Able to Sue its Owner?," *New York Times* (23 April 2014), available online at URL = http://www.nytimes.com/2014/04/27/magazine/the-rights-of-man-and-beast.html?emc=eta1&_r=0>.

[257] See Hanna, *Kantian Ethics and Human Existence*, ch. 4.

[258] See also, for example, Steward, *A Metaphysics for Freedom*.

[259] Directed by R. Wise (1951).

[260] Directed by S. Spielberg (1982).

[261] See, for example, A. Danto, *Analytic Philosophy of Action* (Cambridge: Cambridge Univ. Press, 1973).

[262] See T.S. Eliot, "The Hollow Men," in T.S. Eliot, *Collected Poems* (London: Faber and Faber, 1974), pp. 89-92; and R. Musil, *The Man Without Qualities*, 3 vols. (London: Pan Books, 1979). Significantly, Eliot and Musil both had advanced degrees in philosophy.

[263] See F. Jackson, "Epiphenomenal Qualia," *Philosophical Quarterly* 32 (1982): 127-136.

[264] L. Wittgenstein, *Tractatus Logico-Philosophicus*, trans.C.K. Ogden (London: Routledge, 1981), pp. 109 and 181.

[265] Wittgenstein, *Tractatus Logico-Philosophicus*, pp. 183, 185, and 187.

[266] Dostoevsky, *The Brothers Karamazov*, vol. 2, p. 743.

[267] See, for example, J.M. Fischer, *Deep Control* (Oxford: Oxford Univ. Press, 2012); Fischer and Ravizza, *Responsibility and Control*; M. McKenna, *Conversation and Responsibility* (Oxford: Oxford Univ. Press, 2012); and P.F. Strawson, "Freedom and Resentment," in Watson (ed.), *Free Will*, pp. 72-93.

[268] See also H. Bok, "Freedom and Practical Reason," in Watson (ed)., *Free Will*, pp. 130-166, at p. 130.

[269] This slightly fussy two-part formulation, "really and truly, and perhaps also nothing but," captures both non-reductive physicalist and reductive physicalist versions of Natural Mechanism.

[270] Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem."

[271] T. Huxley, "On the Hypothesis That Animals are Automata, and Its History," in D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings* (New York: Oxford Univ. Press, 2002), pp. 24-30, at pp. 29-30.

[272] I am leaving it open here whether this inconsistency is logical or metaphysical, and if metaphysical, whether it is weakly metaphysical (logical or analytic) or strongly metaphysical (i.e., non-logical or synthetic) inconsistency.

[273] Many thanks to Kristin Mickelson for emphatically making this point to me in conversation and in her PhD dissertation, *Free Will Fundamentals: Agency, Determinism, and (In)Compatibility* (Univ. of Colorado at Boulder, 2012). The point is that Soft Determinism is not metaphysically cheap. It takes *substantive metaphysical work*, and more specifically, substantive metaphysical work in *the philosophy of mind*, to explain it.

[274] Like most contemporary philosophers of free will and agency, Fischer does not distinguish between deep and shallow moral responsibility, and focuses exclusively on shallow responsibility.

[275] See J.M. Fischer, "Frankfurt-type Examples and Semi-Compatibilism," in Kane (ed.), *The Oxford Handbook of Free Will*, pp. 281-308; and J.M. Fischer, *My Way* (New York: Oxford Univ. Press, 2006).

[276] See Pereboom, *Living Without Free Will*; and Pereboom *Free Will, Agency, and Meaning in Life*.

[277] See M. Vargas, "Revisionism," in Fischer, Kane, Pereboom, and Vargas, *Four Views on Free Will*, ch. 4, pp. 126-165.

[278] See, for example, the various views featured in Kane (ed.), *The Oxford Handbook of Free Will*, esp. parts VI and VII. And see also M. Balaguer, *Free Will as an Open Scientific Problem* (Cambridge, MA: MIT Press, 2010); D. Hodgson, *Rationality + Consciousness = Free Will* (Oxford: Oxford Univ. Press, 2012); Ismael, *How Physics Makes Us Free*; E.J. Lowe, *Personal Agency* (Oxford: Oxford Univ. Press, 2008); and Steward, *A Metaphysics for Freedom*.

[279] Griffin, *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem*, p. 171.

[280] See, for example, Kim, *Physicalism, or Something Near Enough.*

[281] See, for example, Prigogine, *The End of Certainty*.

[282] See, for example, Boolos, and Jeffrey, *Computability and Logic*.

[283] See Hanna, *Kant, Science, and Human Nature*, ch. 6; R. Hanna, "Logic, Mathematics, and the Mind: A Critical Study of Richard Tieszen's *Phenomenology, Logic, and the Philosophy of Mathematics*," *Notre Dame Journal of Formal Logic* 50 (2009): 339-361; and *Cognition, Content, and the A Priori*, esp. chs. 6-8. For an opposing view, Maddy, *Second Philosophy*, part IV.

[284] A. Schopenhauer, *On the Fourfold Root of the Principle of Sufficient Reason*, trans. E.F.J. Payne (New York: Open Court, 1974), §42. See also B. O'Shaughnessy, *The Will*, 2 vols. (Cambridge: Cambridge Univ. Press, 1980); B. O'Shaughnessy, *Consciousness and the World* (Oxford: Oxford Univ. Press, 2000); Griffin, *Unsnarling the World-Knot*; and H. Putnam, *The Threefold Cord: Mind, Body, and World* (New York: Columbia Univ. Press, 2001).

[285] Schopenhauer, *The World as Will and Representation*, vol. 1, §20, p. 106.

[286] See, for example, Prigogine, *The End of Certainty*, chs. 3-4.

[287] H. Kleist, "On the Marionette Theater," trans. I. Parry, available online at URL = <http://southerncrossreview.org/9/kleist.htm>.

[288] See, for example, P. Van Inwagen, "The Incompatibility of Free Will and Determinism," *Philosophical Studies* 27 (1975): 185-199; P. Van Inwagen, *An Essay on Free Will* (Oxford: Clarendon/Oxford Univ. Press, 1983); and P. Van Inwagen, "Free Will Remains a Mystery," in Kane (ed.), *The Oxford Handbook of Free Will*, pp. 158-177. See also T. Kapitan, "A Master Argument for Incompatibilism," in Kane (ed.), *The Oxford Handbook of Free Will*, pp. 127-157.

[289] See, for example, Kane (ed.), *The Oxford Handbook of Free Will*, chs. 8-11.

[290] See, for example, M. Fara, "Masked Abilities and Compatibilism," *Mind* 117 (2008): 843-865.

[291] *See, for example, D. Lewis, "Are We Free to Break the Laws?," in Watson (ed).*, Free Will, *pp. 122-129. See also H. Beebee, "Local Miracle Compatibilism,"* Noûs *37 (2003): 258–277; P. Graham, "A Defense of Local Miracle Compatibilism,"* Philosophical Studies *140 (2008): 65-82; and S. Oakley, "Defending Lewis's Local Miracle Compatibilism,"* Philosophical Studies *138 (2006): 337-349.*

[292] W. Holliday, "Freedom and the Fixity of the Past," *Philosophical Review* 121 (2012): 179-207.

[293] N. Tognazzini and J.M. Fischer, "Incompatibilism and the Past," in J. Keller (ed.), *Being, Freedom, and Method: Themes from van Inwagen* (Oxford: Oxford Univ. Press, 2017), pp. 140-148.

[294] Tognazzini and Fischer, "Incompatibilism and the Past."

[295] See H. Frankfurt, "Alternate Possibilities and Moral Responsibility" in Frankfurt, *The Importance of What We Care About*, pp. 1-10; and Frankfurt, "Freedom of the Will and the Concept of a Person." See also N. Levy and M. McKenna, "Recent Work on Free Will and Moral Responsibility," *Philosophy Compass* 4 (2009): 96-133.

[296] See, for example, R. Kane, "Some Neglected Pathways in the Free Will Labyrinth," in Kane (ed.), *The Oxford Handbook of Free Will*, pp. 406-437.

[297] See, for example, R. Adams, "Involuntary Sins," *Philosophical Review*, 94 (1985): 3–31.

[298] See, for example, Fischer, *Deep Control*; Fischer, "Frankfurt-type Examples and Semi-Compatibilism"; Fischer, *My Way*; Fischer and Ravizza, *Responsibility and Control*; McKenna, *Conversation and Responsibility*; and Strawson, "Freedom and Resentment."

[299] See, for example, J.J.C. Smart, "Free Will, Praise, and Blame," *Mind*, 70 (1963): 291–306. The ordinary concept of moral responsibility doesn't seem to track *either* the concept of deep moral responsibility *or* the concept of shallow moral responsibility uniquely so, correspondingly, it seems best interpreted as a more-or-less confused mixture of both.

[300] See Hanna and Maiese, *Embodied Minds in Action*, chs. 1-2; and Hanna, *Kantian Ethics and Human Existence*, ch. 1.

[301] See Kim, "Mechanism, Purpose, and Explanatory Exclusion," in J. Kim, *Supervenience and Mind*, pp. 237-264, esp. at 250.

[302] For a thorough critical discussion of the crucial ambiguities, see Dolch, "A Defense and Interpretation of the Causal Closure of the Physical."

[303] This qualification is crucial to CCP. Consistency yields the criterion of physicality, but entailment or necessitation yields natural mechanization.

[304] This clause rules out *non-standard causal overdetermination*.

[305] This assumption, aka "Funda-Mentalism," is essential for generating The Exclusion Problem. So The Essential Embodiment Theory has no causal-explanatory exclusion worry, and correspondingly, Natural Libertarianism has no exclusion worry either. See Hanna and Maiese, *Embodied Minds in Action*, chs. 6-8.

[306] Kim's own conception of physicality is significantly narrower than mine, arguably begs the the question in favor of reductive physicalism, and above all is clearly committed to a highly questionable assumption that Maiese and I call "Fundamentalism": fundamental physical properties are essentially non-mental, and no substance can have two essences. Fundamentalism is the flip side of Funda-Mentalism, and both are ultimately Cartesian in provenance. See Hanna and Maiese, *Embodied Minds in Action*, section 7.1.

[307] See Kim, "Mechanism, Purpose, and Explanatory Exclusion,"; J. Kim, *Mind in a Physical World* (Cambridge: MIT, 1998), esp. ch. 2; J. Kim, *The Philosophy of Mind* (2nd edn.; Cambridge, MA: Westview Press, 2006), ch. 7; and Kim, *Physicalism, or Something Near Enough* (Princeton, NJ: Princeton Univ. Press 2007). This extension of Kim's causal-exclusion argument to a new and powerful argument for Incompatibilism was first pointed out to me by Kristin Mickelson. For a non-dualist, non-materialist response to Kim's causal-exclusion argument, see Hanna and Maiese, *Embodied Minds in Action*, chs. 7-8.

[308] J. Kim, *Physicalism, or Something Near Enough*, pp. 70-71.

[309] In-the-Zone Compatibilism is also the philosophical brain child of Kristin Mickelson. I'm grateful to her for getting

me to me see the intelligibility of this ingenious view.

[310] See R.E. Hobart, "Free Will as Involving Determinism and Inconceivable Without It," *Mind*, 43 (1934): 1-27.

[311] See, for example, D.F.S. Scott, "Heinrich von Kleist's Kant Crisis," *Modern Language Review* 42 (1947): 474-484; and J. Philips, *The Equivocation of Reason: Kleist Reading Kant* (Stanford, CA: Stanford Univ. Press, 2007).

[312] (Dir. G. Romero, 1968).

[313] See, for example, Hanna, *Kantian Ethics and Human Existence*.

[314] See, for example, Hanna, *Kant, Agnosticism, and Anarchism*, esp. part 3; see also A. Vitale, *The End of Policing* (London: Verso, 2017).

[315] See also Wolf, "Responsibility, Moral and Otherwise."

[316] See D. Pereboom, "Robust Nonreductive Materialism," *Journal of Philosophy* 99 (2002): 499-531; and also D. Pereboom, *Consciousness and the Prospects of Physicalism* (Oxford: Oxford Univ. Press, 2011).

[317] See Hanna and Maiese, *Embodied Minds in Action*, esp. chs. 6-8.

[318] See also Steward, *A Metaphysics for Freedom*, chs. 1-2, where she presents similar arguments for what she calls *Agency Incompatibilism*.

[319] Classical compatibilists and soft determinists typically plump for shallow moral responsibility over deep moral responsibility. But this is not strictly required for their view. See, for example, Frankfurt, *The Importance of What We Care About*. And for a different, non-Frankfurtian version of deep (non-)moral responsibility within the compatibilist/soft determinist tradition, see S. Wolf, "Sanity and the Metaphysics of Responsibility," in F.D. Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (Cambridge: Cambridge Univ. Press, 1987), pp. 46-62. My own view, as per section 4.5 above, is that deep (non-)moral responsibility is locally incompatible with Natural Mechanism. So the compatibilists/soft determinists who favor deep (non-)moral responsibility are in fact committed to an implicit inconsistency, and are in this dilemma: either give up deep (non-)moral responsibility and retain consistency with compatibilism/soft determinism, or else hold onto deep (non-)moral responsibility and go over to Natural Libertarianism. Obviously, I think that they should opt for the second alternative.

[320] This constraint does not affect Hard Incompatibilism, since it is designed specifically to match our best contemporary physical theories, whether deterministic or indeterministic.

[321] See also Hobart, "Free Will as Involving Determinism and Inconceivable Without It."

[322] See notes 307-308 above.

[323] See E. Gettier, "Is Justfied True Belief Knowledge?," *Analysis* 23 (1963): 121-123.

[324] See, for example, D. Pritchard, "Anti-Luck Virtue Epistemology," *Journal of Philosophy* 109 (2012): 247-179. Pritchard also persuasively argues that in addition to the satisfaction of an *externalist* anti-luck condition, a further *virtue-epistemic* "cognitive abilities" condition must also be satisfied in order for justified true belief to count as authentic knowledge. In the free will case, the analogue of the epistemic "abilities" component is built into the capacity for self-commitment to a live option, or Kierkegaardian Either/Or. My own analysis of knowledge, presented in chapter 1 of *Cognition, Content, and the A Priori*, also calls for a third condition to be satisfied in order for justified true belief to count as authentic knowledge, namely an *internalist* "intrinsically-compelling-conscious-evidence" or "self-evidence" condition. In the free will case, the analogue of the epistemic "self-evidence" condition is built into the ownership condition.

[325] See also Steward, *A Metaphysics for Freedom*, ch. 6.

[326] By this, I mean that my complete, finite, and unique far-from-equilibrium, spatiotemporally asymmetric, complex, self-organizing, organismic, finegrainedly normatively attuned thermodynamic life as a rational human animal or real human person up to and including that very moment is *completely* replicated, right down to the most *hyper*-finegrained specific characters of my current vital processes and conscious experiences, and, as a direct consequence, that the complete set of special "one-off" or "one-time-only" causal laws that inherently govern my life up to that very point is also exactly the same as in the actual world.

[327] See, for example, F. Nietzsche, *The Gay Science*, in *The Portable Nietzsche*, trans. W. Kaufmann (Harmondsworth, Middlesex: Penguin, 1983), pp. 101-102, §341.

[328] Those (no doubt) few of us who are existential Kantian cosmopolitan social anarchists, do not accept that the State or State-like institutions, as such, are rationally or morally justified, and also believe that the very idea of political authority as an independent source of obligation, permission, and impermissibility, detached from universal moral principles, is a deeply dangerous philosophical, cognitive, and cultural illusion. See note 11 above. Hence, strictly speaking, it is *impossible* to live fully autonomously in *any* State or State-like institution, in *any* historical context up to today, contemporary Canada included. But perhaps in the future!, somewhere, somehow. Or at least one can rationally hope for this. So it's worth wholeheartedly pursuing it and trying to construct at least its beginnings. All that's packed into what I call the thesis of *radical enlightenment* or *heavy-duty enlightenment*. See Hanna, *Kant, Agnosticism, and Anarchism*, parts 2-3; and Hanna, "Exiting the State and Debunking the State of Nature," THE RATIONAL HUMAN CONDITION, Vol. 1, essay 2.1. But I will bracket full-strength radical enlightenment/heavy-duty enlightenment for the sake of the analogy-based metaphysical point I want to make here. In any case, *obviously*, some real-world States are significantly *more* enlightened and emancipated, and therefore *better* places for people to live, in certain basic respects, than others.

[329] Honesty also requires me to admit that *some* of my Hanna forebears stayed in the USA, and that one of them, John Hanna, was a pall bearer in Lincoln's funeral, and later an Illinois congressman.

[330] See, for example, Chisholm, "Human Freedom and the Self"; Clarke, "Agent Causation and Event Causation in the Production of Free Action"; O'Connor, *Persons and Causes*; and Lowe, *Personal Agency*.
On some interpretations, Kant is a defender of Classical classical-agent-causal Libertarianism; see, for example, Watkins, *Kant and the Metaphysics of Causality*. And on some interpretations, Steward is a Classical Libertarian who also defends a non-classical-agent-causal theory; see, for example, Y. Cohen, Agential Settling Requires a Conscious Intention," *Journal of Cognition and Neuroethics* 3 (2015): 139-155; and J. Runyan, "Events, Agents, and Settling Whether and How One Intervenes," *Philosophical Studies* 173 (2016): 1629-1646. In my opinion, however, because of their views about anti-mechanism and their direct appeals to embodied, animal agency, *neithe*r Kant *nor* Steward is a defender of Classical Libertarianism, whether classical-agent-causal or non-classical-agent-causal. As I read them, Kant and Steward are both saying that intentional agents are immanently-structured life-processes of a certain special kind, *not* Cartesian substances. This makes them neo-Aristotelian hylomorphists, and in effect, Natural Libertarians.

[331] See, for example, C. Ginet, *On Action* (Cambridge: Cambridge Univ. Press, 1990).

[332] See, for example, R. Kane, *The Significance of Free Will* (Oxford: Oxford Univ. Press, 1996); and Hodgson, *Rationality + Consciousness = Free Will*.

[333] Kim, *Physicalism, or Something Near Enough*, ch. 3.

[334] See, for example, C. List and P. Menzies, "Nonreductive Physicalism and the Limits of the Exclusion Principle," *Journal of Philosophy* 106 (2009): 475-502. See also Hanna and Maiese, *Embodied Minds in Action*, pp. 292-294.

[335] See also, for example, P. Van Inwagen, "Free Will Remains a Mystery," in Kane (ed.), *The Oxford Handbook of Free Will*, pp. 158-177.

[336] This is the basic problem with Kane's and Hodgson's otherwise very interesting versions of causal indeterminism—see note 332 above. My own diagnosis of what has gone wrong here is that they confuse *naturally mechanistic indeterminism* (i.e., natural chance or randomness, according to probabilistic or statistical laws) with *non-determinism*-plus-*non-indeterminism* (i.e., natural purposiveness or natural teleology, which is perfectly coherent with agential existence and rational teleology) since both want (i) to deny determinism at the agential source, yet both also want (ii) to rule out causal randomness or agency-undermining luck at the agential source. Nevertheless, naturally mechanistic indeterminism makes the conjunction of (i) and (ii) inconsistent. Only naturally purposive or naturally teleological non-determinism at the source of agency will make them compatible.

[337] See Strawson, "The Impossibility of Moral Responsibility," in Watson (ed.), *Free Will*, pp. 212-228.

[338] Strawson doesn't use this terminology; but if the Basic Argument argument were sound, it would be sufficient to undermine deep moral responsibility in the sense I mean—and also sufficient to undermine deep *non*-moral responsibility too.

[339] J.-P. Sartre, *Being and Nothingness*, trans. H. Barnes (New York: Philosophical Library, 1956), p. 724. Kant calls this same phenomenon "the natural dialectic of pure reason." The basic idea is that we cannot help wanting to be *completely metaphysically grounded*, and also *completely epistemically and practically justified*, in everything we believe and do, and this leads to metaphysical and epistemic disaster in theoretical philosophy. Kant, however, has a finer appreciation than Sartre of the further fact that it is really possible, at the end of the day, in metaphysics, epistemology, and moral philosophy alike, to *accept* our own finitude and to *know* our own limits, in order to pursue the highest good, at least to some salient, life-changing extent. And in this way, Kant was a better Existentialist than Sartre was.

[340] J.M. Fischer, "The Cards that are Dealt You," *Journal of Ethics* 10 (2006): 107-129. Hodgson puts an equally interesting but also, I think, equally misguided *indeterministic* spin on the agency-as-card-playing analogy in *Rationality + Consciousness = Free Will*, pp. 165-166.

[341] Many thanks to Lark Fleming for making this very good point in conversation and also in her MA thesis, "Freedom and Life" (Univ. of Colorado at Boulder, 2008).

[342] See, for example, P. Pettit, *A Theory of Freedom* (Oxford: Oxford Univ. Press, 2001).

[343] Kierkegaard, "Either/Or, A Fragment of Life," p.151.

[344] See Hanna and Maiese, *Embodied Minds in Action*, section 4.2.

[345] See, for example, P. Feyerabend, "Mental Events and the Brain," *Journal of Philosophy* 60 (1963): 295-296; and P. Churchland, "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 78 (1981): 67-90.

[346] See, for example, Churchland, "Eliminative Materialism and the Propositional Attitudes."

[347] Fodor, "Making Mind Matter More," p. 156; and R.E.M., "It's the End of the World as We Know It (and I Feel Fine)," from *Document* (1987), lyrics by M. Stipe.

[348] See, for example, G.E. Moore, "A Defence of Common Sense," "Proof of an External World," and "Certainty," all in G.E. Moore, *G.E. Moore: Selected Writings* (London: Routledge, 1993), pp. 106-133, 147-170, and 171-196; see also Lowe, *Personal Agency*, pp. 197-198.

[349] See Pereboom, *Living Without Free Will*; and Pereboom, *Free Will, Agency, and Meaning in Life*.

[350] See, for example, D. Pereboom, "The Phenomenology of Agency and Deterministic Agent-Causation," in M. Altman and H. Gruenig (eds.), *Horizons of Authenticity in Phenomenology, Existentialism, and Moral Psychology: Essays in Honor of Charles Guignon* (New York: Springer, 2015), pp. 277-294.

[351] Kierkegaard, "Either/Or, A Fragment of Life," p. 72.

[352] Frankfurt, "Freedom of the Will and the Concept of a Person," p. 25, underlining added.

[353] Kane, *The Significance of Free Will*, p. 35.

[354] See Hanna, *Cognition, Content, and A Priori*, ch. 7.

[355] See, for example, the various free will theories surveyed in Kane, *A Contemporary Introduction to Free Will*. Kane very usefully ends the book with a survey of five aspects or types of freedom of the will: (i) self-realization, (ii) reflective self-control, (iii) self-perfection, (iv) self-determination, and (v) self-formation. All five of these aspects or types of freedom are accounted-for within the ordered, three-leveled metaphysical structure postulated by Natural Libertarianism.

[356] See also Hanna and Maiese, *Embodied Minds in Action*, esp. section 6.1.

[357] More specifically, my claim is that real causation is a relation of non-logical, "strong metaphysical," synthetic a priori necessitation, indexed to the egocentrically-centered actual world, that obtains only under special ceteris paribus contextual conditions, fully consistent with all deterministic or indeterministic general causal laws of nature, especially including the Conservation Laws, together with all settled quantity-of-matter-and/or-energy facts about the past, especially including The Big Bang, *but not necessarily entailed or otherwise necessitated by that base*, for example, in cases in which the cause is really causally spontaneous, hence naturally purposive or teleological, hence non-deterministic but also non-indeterministic. Obviously, I cannot adequately develop and defend this theory here, only state it and indicate its explanatory advantages over other approaches. But the crucial point for the present purposes is that this approach to causation gets significantly between (i) classical modal theories of causation grounded on hypothetical logical-conceptual, "weak metaphysical," or analytic necessity, on the one hand, which are characteristically too strong, and (ii) non-modal theories, on the other, which are characteristically too weak. The standard dichotomy between (i)-type logico-conceptual-analysis-grounded theories, following on from Logical Empiricism, and (ii)-type scientific-naturalism-grounded theories, following on from post-Empiricist Quineanism or neo-Quineanism, presuppose modal monism on both sides of the dichotomy. But I want to reject modal monism, in favor of a contemporary Kantian version of modal dualism—see, for example, Hanna, *Cognition, Content, and the A Priori*, ch. 4—and, correspondingly, claim that the standard modal/non-modal dichotomy in the theory of causation is a false one that can be overcome by the sort of theory I'm proposing.

[358] As well as being causal-nomological *underdeterminers* of them, as per the second condition on spontaneity.

[359] See, for example, Pettit, *A Theory of Freedom*.

[360] See B. O'Shaughnessy, "Trying (as the Mental 'Pineal Gland'), " in A. Mele (ed.), *The Philosophy of Action* (Oxford: Oxford Univ. Press, 1997), pp. 53-74; and Hanna and Maiese, *Embodied Minds in Action*, chs. 3-5 and section 8.3.

[361] See Sartre, *Being and Nothingness*, part 4.

[362] Frankfurt, "Alternate Possibilities and Moral Responsibility," p. 1.

[363] Frankfurt, "Alternate Possibilities and Moral Responsibility," pp. 6-9, underlining added.

[364] See Frankfurt, "Freedom of the Will and the Concept of a Person," pp. 20-25.

[365] D. Widerker, "Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities," in Watson (ed.), *Free Will*, pp. 177-189.

[366] See, for example, D. Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (Cambridge: MIT Press, 1984), p. 133.

[367] J.M. Fischer, "The Frankfurt Cases: The Moral of the Stories," *Philosophical Review* 119 (2010): 315-336. A similar point is made by Lowe in *Personal Agency*, pp. 156-157 and 196-197.

[368] For a detailed analysis of the nature and structure of consciousness like ours, see Hanna and Maiese, *Embodied Minds in Action*, chs. 1-2.

[369] Analogously, I am committed to strong metaphysical disjunctivism about sense perception. See Hanna, *Cognition, Content, and the A Priori*, ch. 3.

[370] See Hanna, *Cognition, Content, and the A Priori*, ch. 2.

[371] On my view, non-human animals and infant humans, etc., who are capable only of free volition, but not free will, are also capable of veridical psychological free volition, as well as non-veridical psychological free volition. And correspondingly there is a weak metaphysical disjunctivist analogue of condition (iii) that holds for psychological free volition, namely:

(iii*) veridical psychological free volition and non-veridical psychological free volition essentially share no intentional or phenomenological content except for whatever it is that actually makes veridical psychological free volition and non-veridical psychological volition indiscriminable for non-rational minded animal agents, even if they accidentally share other psychological or non-psychological properties, such that the animal agent is either in one kind of state or the other, never both.

Obviously, (iii*) entails that veridical psychologically free volition and non-veridical psychologically free volition are indiscriminable for non-rational minded animal agents. Hence there is no analogue of condition (iv)

for them.

[372] C. Bass, "By Reason of Insanity," *Yale Alumni Magazine* (May/June 2015): 48-53, at p. 49, available online at URL = <http://www.yalealumnimagazine.com/articles/4076/by-reason-of-insanity>.

[373] See Frankfurt, "Freedom of the Will and the Concept of a Person."

[374] See note 372 above, and also, for example, C. Frith, *The Cognitive Neuropsychology of Schizophrenia* (Hillsdale, NJ: Erlbaum, 1992); C.S. Mellor, "First Rank Symptoms of Schizophrenia," *British Journal of Psychiatry* 117 (1970): 15-23; and S.A. Spence, "Free Will in the Light of Psychiatry," *Philosophy, Psychiatry, and Psychology* 3 (1996): 75-90.

[375] Frankfurt, "Alternate Possibilities and Moral Responsibility"; and Frankfurt, "Freedom of the Will and the Concept of a Person."

[376] Harris, *Free Will*, p. 5.

[377] See, for example, N. Arpaly, *Merit, Meaning, and Human Bondage* (Princeton, NJ: Princeton Univ. Press, 2007).

[378] See Kant, *Groundwork of the Metaphysics of Morals*, sections II-III (*GMM* 4: 406-463); and Kant, *Critique of Practical Reason*, ch. I, (*CPrR* 5: 19-57).

[379] See Frankfurt, "Freedom of the Will and the Concept of a Person"; and Frankfurt, "Identification and Wholeheartedness." Frankfurt holds that decisive identification is a necessary condition of (deep) moral responsibility, but that seems clearly false. Surely I can be morally responsible for choosing or doing *X* even if I am significantly emotionally conflicted about the proper ordering of my desires? Otherwise, merely being significantly confused about my real feelings or upset about the overall coherence of my inner life would be a sufficient condition for undermining moral responsibility. And that can't be right, in that presumably most difficult choices and difficult decisions for most real persons at most times are made under conditions of significant emotional conflictedness and stress. Significant emotional conflictedness, to be sure, undermines Kantian principled authenticity, or the realization of autonomy; but only *overwhelming* emotional conflictedness, which entails psychological or metaphysical *unfreedom*, would undermine moral responsibility.

[380] See Frankfurt, "Identification and Wholeheartedness"; and Frankfurt, *The Reasons of Love*.

[381] This is an essential feature of Sartrean freedom. See, for example, S. Crowell, "Sartre's Existentialism and the Nature of Consciousness," in S. Crowell (ed.), *The Cambridge Companion to Existentialism* (Cambridge: Cambridge Univ. Press, 2012), pp. 199-226.

[382] This account of reasons for action is also developed in detail in Hanna and Maiese, *Embodied Minds in Action*, ch. 3.

[383] See, for example, R. Solnit, *A Paradise Built in Hell: The Extraordinary Communities That Arise in Disaster* (New York: Penguin, 2009); and L. MacFarquhar, *Strangers Drowning: Impossible Idealism, Drastic Choices, and the Urge to Help* (New York: Penguin, 2016). These two books, in effect, make a complementary pair: MacFarquhar's book shows that human altruism is not only psychologically really possible, but also fully empirically confirmed as a special psychological profile (so psychological egoism is false); and Solnit's book shows that the human capacity for altruism has not only been repeatedly exemplified in ordinary people under extraordinary contextual and historical conditions, but also is thoroughly morally admirable (so ethical egoism is also false).

[384] Wittgenstein, *Philosophical Investigations*, §109, p. 97e.

[385] (Dir. D. Siegel, 1956).

[386] I borrow the evocative label "sinner-saint" from Lillian Hellman's "Introduction," in D. Hammett, *The Big Knockover* (London: Orion Books, 2005), pp. v-xxii, where she talks about parallels between some of the central characters in Dostoevsky's novels and Hammett himself. The very idea of Dostoevskian sinner-sainthood, in turn, also captures the apparently paradoxical truth, noted in chapter 3, that near-Satanic evil, in certain motivational respects, is fundamentally closer to principled authenticity than either non-evil moral badness or banal evil.

Still, no wonder Hammett was deeply hurt when Hellman called him a "Dostoevskian sinner-saint." And Hellman's role in Hammett's genuinely tragic life was not entirely without its own significant measure of Dostoevskian sinner-sainthood, as per, for example, Grushenka in *The Brothers Karamazov*.

[387] Frankfurt, "Freedom of the Will and the Concept of a Person," p. 12.

[388] E. Olson, *The Human Animal* (Oxford: Oxford Univ. Press, 1997), p. 124.

[389] This idea goes back to Hobbes. See, for example, R. Hanna, "Persons and Personation in Hobbes's *Leviathan*," *Southern Journal of Philosophy* 21 (1983): 177-191.

[390] See, for example, L. Baker, *Persons and Bodies: A Constitution View* (Cambridge: Cambridge Univ. Press, 2000); R. Chisholm, *Person and Object* (La Salle, IL: Open Court, 1976); Frankfurt, "Freedom of the Will and the Concept of a Person"; H. Hudson, *A Materialist Metaphysics of the Human Person* (Ithaca, NY: Cornell University Press, 2001); E.J. Lowe, *Subjects of Experience* (Cambridge: Cambridge Univ. Press, 1996); E. Olson, *What Are We? A Study in Personal Ontology* (Oxford: Oxford Univ. Press, 2007); P.F. Strawson, *Individuals* (London: Methuen, 1959), ch. 3; D. Wiggins, *Sameness and Substance* (Oxford: Blackwell, 1980); K. Wilkes, *Real People* (Oxford: Clarendon/Oxford Univ. Press, 1988); and Wollheim, *The Thread of Life*.

[391] See, for example, J. Dancy (ed.), *Reading Parfit* (Oxford: Blackwell, 1997); R. Martin and J. Barresi (eds.), *Personal Identity* (Oxford: Blackwell, 2003); H. Noonan, *Personal Identity* (2nd edn, London: Routledge, 2003);

D. Parfit, *Reasons and Persons* (Oxford: Oxford Univ. Press, 1984); J. Perry (ed.), *Personal Identity* (Berkeley, CA: Univ. of California Press, 1975); A. Rorty (ed.), *The Identities of Persons* (Berkeley, CA: Univ. of California Press, 1976); S. Shoemaker and R. Swinburne, *Personal Identity* (Oxford: Blackwell, 1984); and P. Unger, *Identity, Consciousness, and Value* (Oxford: Oxford Univ. Press, 1990).

[392] See also Wollheim, *The Thread of Life*. Despite the many important parallels between Wollheim's account and mine, there are also four important differences. First, although Wollheim very rightly (and brilliantly) employs a phenomenological method to establish his claims, he provides no background metaphysics of essential embodiment to vindicate his phenomenology of embodiment, whereas I do provide this; second, Wollheim's background account of the underlying mentalistic structure and psychodynamics of the real person's life is basically Freudian, whereas mine is basically Kantian and Existentialist; third, Wollheim does not explicitly connect his account with a metaphysics of free agency, whereas my account both constitutively presupposes and is essentially complementary to Natural Libertarianism; and fourth, while, like my account, Wollheim's life-oriented, real person account is explicitly directed *against* the then-standard approaches, in the early 1980s, to the problem of personal identity—although it preceded Animalism, which was a metaphysical child of the 90s, and therefore, naturally,Wollheim doesn't mention it—I also use the life-oriented, real-person account to work out a new metaphysical approach to the identity issue, and correspondingly formulate a new criterion of personal identity.

[393] For more on the concept of death, and a theory of the morality of our own deaths, see Hanna, *Kantian Ethics and Human Existence*, ch. 6.

[394] Michelle Maiese, in *Embodied Selves and Divided Minds* (Oxford: Oxford Univ. Press, 2015), defends a view about the nature of selves that she also calls "Minded Animalism." Obviously, as co-authors of *Embodied Minds in Action*, our views are not likely to be *very* dissimilar, and have indeed significantly mutually influenced one another. The principal difference between my version of Minded Animalism and hers, however, is that my conception of real persons closely tracks the capacity for 2D rationality, even if it is only lower-level/instrumental or minimal rationality, whereas she allows for persons/selves who are wholly non-rational. As to the crucial borderline case of insanity, my version of Minded Animalism says that those insane human minded animals who are permanently incapable of any degree of rationality, free will, or practical agency, are *non*-persons—although even the permanently, deeply insane do indeed retain "minimal selfhood," in the sense that they are still individual conscious embodied subjects. By contrast, the temporarily, shallowly, or more generally curably insane are, precisely to the extent that their mental illnesses are temporary or shallow, and curable, also correspondingly and to that extent capable of rationality, also capable of free will and practical agency. Hence they are real persons. As Susan Wolf correctly points out in "Responsibility, Moral and Otherwise," permanently or temporarily insane human beings can be deeply non-morally responsible for various free choices and acts, even if they are not deeply morally responsible for them. Indeed, there is a large range of cases in which human beings have free volition, and thus deep non-moral responsibility, like many non-human animals, but not free will or deep moral responsibility. See section 4.1 above.

[395] This sense is particularly brillantly and even movingly conveyed by Wollheim's *Thread of Life*, against a Freudian backdrop. Relatedly, although not in a Freudian framework, I will present a "Shakespearian" evidential/phenomenological argument for Minded Animalism, that I call *The Ecce Homo Argument*, in section 6.1 below.

[396] On the deflationary side, there are also skeptical and/or social constructivist views of persons and selves, that reject the real existence of persons or selves, or at least make persons and selves strongly supervenient on historically contingent social communities and sociocultural processes. As to the fully skeptical view, it is hard to see how this view could be *rationally* defended, without covertly presupposing a real person who is attempting to justify her beliefs by offering reasons for them, and therefore without rational self-contradiction or self-stultification. As to the strong supervenience view, I don't deny that the *concept* of the person or self is significantly determined by historically contingent social communities and sociocultural processes. But since a real person or self is categorically more than the *concept* of a person or self, it is simply a non sequitur to infer from the social construction of concepts, to the social construction of the things described by those concepts.

[397] Indeed, there is very good reason to believe that an obsession with the moral-political goal of equal treatment, and with reducing relatively minor inequalities between classes of relatively well-off people, is deeply morally misguided, For it is in fact a violation of respect for the human dignity of the poorest members of society, and more generally a violation of respect for the human dignity of oppressed people of all kinds, everywhere, to whom we always morally owe *enough* to satisfy the moral requirements of respect for their human dignity. See, for example, H. Frankfurt, *On Inequality* (Princeton, NJ: Princeton Univ. Press, 2015).

[398] Nevertheless, a real-life 20th century American president, Ronald Reagan, did actually star in a movie about a chimpanzee, called *Bedtime for Bonzo*.

If, however, I am right that the very idea of the State is rationally unjustified and immoral (see Hanna, *Kant, Agnosticism, and Anarchism*, esp. part 2), then it is not altogether surprising that a former second-rate Hollywood actor, a president of the Screen Actors' Guild during the McCarthy era, and an assiduous informant for HUAC, would eventually get to be both Governor of California and also a two-term US President. As for the current so-called US President, Donald Trump, I'll leave the relevant extension of the *Bedtime for Bonzo* analogy to the

imagination of the reader.

[399] See also, for example, P. Snowdon, "Persons, Animals, and Ourselves," in C. Gill (ed.), *The Person and the Human Mind* (Oxford: Oxford Univ. Press, 1990), pp. 83-108.

[400] J.-P. Sartre, "Existentialism is a Humanism," in Cahn and Markie (eds), *Ethics: History, Theory, and Contemporary Issues*, pp. 406-412, at p. 412.

[401] Nevertheless, in another philosophical context, i.e., the context of moral philosophy, this is indeed a problematic implication. See Hanna, *Kantian Ethics and Human Existence*, ch. 3.

[402] See, for example, E. Olson, "Animalism and the Corpse Problem,"*Australasian Journal of Philosophy* 82 (2004): 265-274, also available online at URL = http://eprints.whiterose.ac.uk/718/1/olsonet3.pdf; D. Hershenov, "Do Dead Bodies Pose a Problem for Biological Approaches to Personal Identity?," *Mind* 114 (2005): 31-59; J. LaPorte, "On Two Reasons for Denying that Bodies Can Outlast Life," *Mind* 118 (2009): 795-801; and D. Hershenov, "Organisms and their Bodies: A Reply to LaPorte," *Mind* 118 (2009): 803-809.

[403] See K. Rowland, "We Are Multitudes," *Aeon* (11 January 2018), available online at URL = <https://aeon.co/essays/microchimerism-how-pregnancy-changes-the-mothers-very-dna>.

[404] A. Bailey, "Animalism," *Philosophy Compass* 10 (2015): 867-883.

[405] For a detailed theory of rational intuitions, a priori justification, and a priori knowledge, see Hanna, *Cognition, Content, and the A Priori*, chs. 6-8.

[406] See note 397 above.

[407] The more-or-less online character of these capacities is of crucial importance for distinguishing between Frankfurtian real persons and Kantian real persons, but not for the more general point I am making here, since although Frankfurtian real persons and Kantian real persons differ importantly in the ways in which their basic capacities are configured and disposed, nevertheless their absolute intrinsic non-denumerable objective value or dignity is exactly the same, precisely because their basic capacities are the same.

[408] See Hanna, *Kantian Ethics and Human Existence*, ch. 2.

[409] This point is shown by Judith Thomson's "loop variant" on the standard Trolley Problem in "The Trolley Problem," in S. Cahn and P. Markie (eds.), *Ethics: History, Theory, and Contemporary Issues* (4th edn.; Oxford: Oxford Univ. Press, 2010), pp. 910-923, and also by D. Parfit's "George-the-gangster" case in his *On What Matters* (Oxford: Oxford Univ. Press, 2011), vol. 1, ch. 9. See also F. Kamm, *Intricate Ethics* (Oxford: Oxford Univ. Press, 2006), ch. 5; and F. Kamm, *Morality, Mortality*, 2 vols. (Oxford: Oxford Univ. Press, 1996), vol. II, chs. 6-7.

[410] See Hanna, *Kantian Ethics and Human Existence*, ch. 5.

[411] This qualification is important, because according to Minded Animalism there are some real persons whose basic capacities for free agency are, as it so happens, *not fully online*, and therefore are to some extent immature, latent, or undeveloped, for example, normal human toddlers and other children, and some species of non-human animals. These are the non-autonomous, lower-level, Frankfurtian real persons, real persons$_f$.

[412] See Hanna and Maiese, *Embodied Minds in Action*, section 5.3.

[413] Maiese and I spell out and defend this theory of the emotions in detail in *Embodied Minds in Action*, ch. 5; she has also worked out and defended this view separately in *Embodiment, Emotion, and Cognition* (London: Palgrave Macmillan, 2011); and I have spelled out and defended the theory of the essentially non-conceptual content of perception in detail in *Cognition, Content, and the A Priori*, ch. 2. For a brief presentation of the latter, see also section 2.4 above.

[414] Wittgenstein, *Philosophical Investigations*, p. 178e.

[415] See, for example, T. Nagel, "Moral Luck," in T. Nagel, *Mortal Questions* (Cambridge: Cambridge Univ. Press, 1979), pp. 24-38; B. Williams, "Moral Luck," in B. Williams, *Moral Luck* (Cambridge: Cambridge Univ. Press, 1981), pp. 20-39; D. Domsky, "There is No Door: Finally Solving the Problem of Moral Luck," *Journal of Philosophy* 101 (2004): 445-464; D. Statman, "Doors, Keys, and Moral Luck: A Reply to Domsky," *Journal of Philosophy* 102 (2005): 422-436; and D. Nelkin, "Moral Luck," *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), E.N. Zalta (ed.), available on line at URL = <http://plato.stanford.edu/archives/win2013/entries/moral-luck/>.

[416] See, for example, J.-P.Sartre, *Existentialism and Human Emotions*, trans. B. Frechtman and H. Barnes (New York: Citadel Press, 1990); and Sartre, *Being and Nothingness*, part 4.

[417] This non-moral obligation to resist inathenticity would hold in cases of, for example, a creative artist who "sells out," as in the basic lyrics of the classic 10cc 1970s pop song: *art for art's sake, money for God's sake!*

[418] See note 403 above.

[419] See, for example, P. Griffiths, "Darwinism, Process Structuralism, and Natural Kinds," *Philosophy of Science* 63 (1996): S1-S9; and P. Griffiths, "Squaring the Circle: Natural Kinds with Historical Essences," in R. Wilson (ed.), *Species: New Interdisciplinary Essays* (Cambridge, MA: MIT Press, 1999), pp. 209-228.

[420] See, for example, J. O. La Mettrie, *L'Homme Machine/Man a Machine* (Chicago, IL: Open Court, 1912), available online at URL = <https://archive.org/details/manmachine00lame>.

[421] See, for example, S. Wolf, "Moral Saints," *Journal of Philosophy* 79 (1982): 419-439. Wolf argues that we should not want to be or strive to be (more like) moral saints. I completely agree. But it does not follow that there cannot be real-world moral saints, or "sinner-saints." Indeed, I think that there really *are* such people, and *many more*

of them than we might at first think. Moreover, I also think that we should all want to be and strive to be more like them—not as *sinner*-saints, of course, but rather as sinner-*saints*.

[422] Autobiographical note: I've been carrying a copy of this poem, cut out of the *New York Times Book Review*, now utterly yellowed, and completely falling apart, but for the Scotch tape that just barely holds it together, in a succession of wallets, since 1977. Why?

[423] See Hanna, *Kantian Ethics and Human Existence*, ch. 4.

[424] J. Hawkins and R. Allen (eds.), *Oxford Encyclopedic English Dictionary* (Oxford: Clarendon/Oxford Univ. Press, 1991), p. 52.

[425] See, for example, C. Allen and M. Bekoff, *Species of Mind* (Cambridge: MIT Press, 1997); M. Bearzi and C. Stanford, *Beautiful Minds: The Parallel Lives of Great Apes and Dolphins* (Cambridge: Harvard Univ. Press, 2008); D.R. Griffin, *Animal Minds* (Chicago: Univ. of Chicago Press, 2001); D.R. Griffin, *Animal Thinking* (Cambridge: Harvard Univ. Press, 1984); and D.R. Griffin, *The Question of Animal Awareness* (New York: Rockefeller Univ. Press, 1976).

[426] See Hanna and Maiese, *Embodied Minds in Action*, esp. chs. 1-2.

[427] See Hanna and Maiese, *Embodied Minds in Action*, esp. chs. 3, 4, and 5.

[428] See H. Putnam, *Reason, Truth, and History* (Cambridge: Cambridge Univ. Press, 1981), ch. 1.

[429] See Hanna and Maiese, *Embodied Minds in Action*, section 8.1.

[430] The relevant set of neurobiological properties alone is not a *sufficient* condition of the existence of a consciousness like ours, however. Instead, the existence of consciousness like ours is *jointly hylomorphically constituted* by relevant mental and neurobiological properties. See Hanna and Maiese, *Embodied Minds in Action*, section 8.1.

[431] Just as in the case of the existence of consciousness like ours, so too the relevant set of neurobiological properties alone is not a sufficient condition of the *specific character* of consciousness like ours. Both the existence and specific character of a consciousness like ours are jointly hylomorphically constituted by relevant mental and neurobiological properties. Again, see Hanna and Maiese, *Embodied Minds in Action*, section 8.1.

[432] Recent cognitive neuroscience shows some healthy signs of gradually shifting from a brain-centered approach to a whole-body-centered approach. See, for example, L. Shapiro (ed.), *The Routledge Handbook of Embodied Cognition* (London: Routledge, 2014).

[433] See, for example, P. Godfrey-Smith, *Other Minds: The Octopus and the Evolution of Intelligent Life* (New York: Collins, 2017); and A. Srinivasan, "The Sucker, the Sucker!," *London Review of Books* 39 (September 2017): 23-25, available online at URL = <https://www.lrb.co.uk/v39/n17/amia-srinivasan/the-sucker-the-sucker?utm_source=newsletter&utm_medium=email&utm_campaign=3917&utm_content=usca_subs>.

[434] See, for example, M. Gershon, *The Second Brain* (New York: HarperCollins, 1998).

[435] See, for example, A. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Avon Books, 1994); A. Damasio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (San Diego, CA: Harcourt, 1999); A. Damasio, *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain* (San Diego, CA: Harcourt, 2003); C. Pert, *Molecules of Emotion* (New York: Scribner, 1997); and J. Prinz, *Gut Reactions: A Perceptual Theory of Emotion* (New York: Oxford, 2004).

[436] See, for example, Shapiro (ed.), *The Routledge Handbook of Embodied Cognition*.

[437] See Hanna and Maiese, *Embodied Minds in Action*, section 1.2; and Hanna, *Cognition, Content, and the A Priori*, section 2.8.

[438] See H. Kuhse and P. Singer, "Individuals, Humans, and Persons: The Issue of Moral Status," in P. Singer, H. Kuhse, S. Buckle, K. Dawson, and P. Kasimba (eds.), *Embryo Experimentation* (Cambridge: Cambridge Univ. Press, 1990), pp. 65-75. Kuhse and Singer hold that some living organism is an individual only if it leaves a corpse behind when it dies. But obviously one way of ending the life of an individual living organism is to consume or otherwise completely destroy its body. Another but much less obvious way to end the life of an individual living organism without leaving a corpse behind is to sever parts of its body such that the severed parts continue to live on their own as distinct human individuals and real persons--see the strange real world case of the Filipino *janus* twins discussed in chapter 7 below.

[439] See also Hanna, *Kantian Ethics and Human Existence*, sections 3.1-3.3.

[440] See D. Boonin, *A Defense of Abortion* (Cambridge: Cambridge Univ. Press, 2003), ch. 3. For an earlier study that puts the emergence of consciousness$_{lo}$ at between 22-26 weeks, see *British Parliamentary Office of Science and Technology Notes* 94 (1997), available at URL = <http://www.parliament.uk/post/pn094.pdf>.

[441] A famous example is the real-world case of Baby Theresa. See J. Rachels and S. Rachels, *The Elements of Moral Philosophy* (6th edn., New York: McGraw-Hill, 2010), pp. 1-5.

[442] See Hanna, *Kantian Ethics and Human Existence*, ch. 3.

[443] T. Nagel, "Brain Bisection and the Unity of Consciousness," in Nagel, *Mortal Questions*, pp. 147-164, at 153-154.

[444] See notes 372 and 374 above.

[445] A famous example is the real-world case of Jodie and Mary. See Rachels and Rachels, *The Elements of Moral Philosophy*, pp. 5-7.

[446] It is possible for the same human individual to be a fully-constituted person at some times in its life (for example, Iris Murdoch at age 30), a neo-person at other times in its life (for example, Iris Murdoch as a 3rd trimester fetus), and a non-person at still other times in its life (for example, Iris Murdoch in the advanced stages of Alzheimer's

disease).

[447] See Hanna and Maiese, *Embodied Minds in Action*, chs. 1-2; and Hanna, "Minding the Body."

[448] See Frankfurt, "Identification and Externality," in Frankfurt, *The Importance of What We Care About*, pp. 58-68; and Frankfurt, "Identification and Wholeheartedness."

[449] See Griffin, *Animal Minds*; Bearzi and Stanford, *Beautiful Minds: The Parallel Lives of Great Apes and Dolphins*; and S. Savage-Rumbaugh and R. Lewin, *Kanzi: The Ape at the Brink of the Human Mind* (New York: Wiley, 1994). Savage-Rumbaugh's research is highly controversial. For an alternative view, see M. Tomasello and J. Call, *Primate Cognition* (New York: Oxford Univ. Press, 1997), esp. pp. 375-379. My own view, which I spell out and defend in *Kantian Ethics and Human Existence*, says that Great apes and perhaps also dolphins are *non-autonomous persons* who are morally equivalent to normal human toddlers and other young children. This in turn suggests an argument strategy for those who seek to extend person-based legal rights to Great apes and dolphins: Since normal human toddlers and other young children *clearly* have real personhood and dignity, and since Great apes and (perhaps also) dolphins possess the same psychological capacities that ground real personhood and dignity, then it follows that Great apes and (perhaps also) dolphins *also* have real personhood and dignity, and therefore *should also* be accorded the same person-based legal rights. See Siebert,"Should a Chimp Be Able to Sue its Owner?"

[450] See section 3.3 above.

[451] See section 3.4 above.

[452] See, for example, D. Hume, *Treatise of Human Nature* (2nd edn., Oxford: Clarendon/Oxford Univ. Press, 1978), books II and III.

[453] See, for example, M. Tomasello, *Why We Cooperate* (Cambridge, MA: MIT Press, 2009); and P. Bloom, "The Moral Life of Babies," *New York Times Magazine* (9 May 2010).

[454] For a theory of the morality of our own deaths, see Hanna, *Kantian Ethics and Human Existence*, ch. 7.

[455] In Frankfurt's terminology, normal human toddlers and other children cannot identify or decide "wholeheartedly." See Frankfurt, "Identification and Wholeheartedness."

[456] Another autobiographical note: Speaking of the abyss, or rather of The Grund as *der Abgrund*, in mid-May 2014, walking home from the Centre Ville of Luxembourg, I saw a teenager showing off to his friends, laughing, standing on the stone parapet of the Montée Clausen, above The Grund. Then he jumped a metre or so across onto another parapet, less than 6 centimeters away from a fall of at least 100 metres straight down into The Grund; then he jumped back onto the Montée bridge, and simply walked away with his friends, still laughing. It was simply unbearable to watch, it made me feel sick to my stomach thinking about it afterwards, and thank the Highest Good that he didn't fall. But it perfectly exemplified the semi-Kantian character of adolescents.

[457] See Hanna, *Rationality and Logic*, esp. chs. 6-7.

[458] See, for example, Hanna, *Kant and the Foundations of Analytic Philosophy*, chs. 3-5; Hanna, *Cognition, Content, and the A Priori*, chs. 5-6; and Hanna, "Kant, the Copernican Devolution, and Real Metaphysics." It's difficult to overstate the philosophical importance of the debate and issues surrounding the analytic-synthetic, a priori-a posteriori, and necessary/contingent truth/falsity distinctions, even despite the sociological fact that virtually all recent and contemporary philosophers find it either "obviously" resolved in favor of some sort of classical, pragmatic, or post-Quinean Empiricism, or else rebarbatively tedious, yet cannot give up *some* sort of implicit commitment to it, via their views about logic and mathematics.

[459] See Hanna, *Kantian Ethics and Human Existence*, ch. 4.

[460] See Hanna, *Kantian Ethics and Human Existence*, chs. 3-4.

[461] See M. Tooley, "Abortion and Infanticide," in M. Cohen, T. Nagel, and T. Scanlon (eds.), *The Rights and Wrongs of Abortion* (Princeton, NJ: Princeton Univ. Press, 1974), pp. 52-84. In effect, even if not explicitly, Tooley sets the bar for personhood at the level of Kantian (2D) rationality. But this is much too high, because it certainly fails to include toddlers, and also possibly even fails to include adolescent, semi-autonomous persons, depending on how one construes the conditions on concept-*possession*. In any case, according to The Real Person Theory, normal third trimester fetuses, infants, toddlers, and adolescents are all real persons too.

[462] See, for example, Tomasello and Call, *Primate Cognition*; and Tomasello, *Why We Cooperate*.

[463] Frankfurt, "Freedom of the Will and the Concept of a Person," p. 17.

[464] Frankfurt, "Identification and Wholeheartedness," p. 176.

[465] Leaving aside for the moment, for simplicity's sake, the intermediate class of semi-Kantian (middle rank) real persons.

[466] D. Parfit, *Reasons and Persons* (Oxford: Clarendon/Oxford Univ. Press, 1984), p. ix.

[467] See section 1.0 above.

[468] Parfit, *Reasons and Persons*, p. ix.

[469] Ibid, p. 192.

[470] Ibid, p. 317.

[471] Ibid, p. 347.

[472] Ibid, pp. 346-347.

[473] The same conclusion is also reached by Frances Kamm, but like Parfit, on non-Kantian grounds. See, for example, Kamm, *Intricate Ethics*.

[474] I agree with Parfit here, that compensation made to adults often cannot make up for the harms inflicted on children, but that is because I think that children are real persons, and subjects of dignity, too. See Hanna, *Kantian Ethics and Human Existence*, chs. 3-4.

[475] Again, I agree with Parfit here, that phenomenology has moral value, and indeed intrinsic moral value; but, for me, *only within the lives of real persons*, and only because that phenomenology is is actually an essentially embodied *bio*phenomenology that is grounded on the deeper fact of real personhood. See sections 6.2 and 6.3 above.

[476] My own view is that punishment, as such, is rationally unjustified and immoral. But *if* punishment *were* rationally justified and morally permissible, then I agree with Parfit that punishment should be morally sensitive to the links between the phenomenological phases of individual lives, but, for me, only within the larger context of the complete structure of a real person's life. See Hanna, *Kantian Ethics and Human Existence*, ch. 4, and Hanna, *Kant, Agnosticism and Anarchism*, part 3.

[477] And yet again, I agree with Parfit here, that the fertilized ovum is neither an individual human being nor a subject of moral value, far less a real person, and that the salient moral properties of the fetus are emergent features of the developing pregnancy. But, for me, this is only because these moral properties are grounded on the deeper fact of real personhood, which emerges with neo-personhood in the third trimester of pregnancy. See Hanna, *Kantian Ethics and Human Existence*, ch. 3.

[478] Parfit, *Reasons and Persons*, p. 202.

[479] See J. Locke, *Essay concerning Human Understanding* (Oxford: Clarendon/Oxford Univ. Press, 1975), book II, ch. xxvii, §§ 9-29, pp. 335-348.

[480] Parfit, *Reasons and Persons*, p. 216.

[481] Locke's actual account of personal identity is much more subtle and interesting—see, for example, M. Ayers, *Locke* (2 vols., London: Routledge: 1991), vol. 2, chs. 22-25.

[482] Parfit, *Reasons and Persons*, p. 216.

[483] See, for example, Kim, *Supervenience and Mind*, esp. part 1; Horgan, "From Supervenience to Superdupervenience: Meeting the Demands of a Material World"; Chalmers, *The Conscious Mind*, chs. 1-3; Kim, *Philosophy of Mind*, chs. 1 and 10; Kim, *Physicalism, or Something Near Enough*; and Hanna and Maiese, *Embodied Minds in Action*, Introduction and section 1.1.

[484] Parfit, *Reasons and Persons*, pp. 216-217.

[485] Ibid, p. 217.

[486] Ibid, p. 201.

[487] Ibid, p. 217.

[488] Apologies for this extra-ugly adverbial neologism.

[489] B. Williams, "The Self and the Future," in B. Williams, *Problems of the Self* (Cambridge: Cambridge Univ. Press, 1973), pp. 46-63.

[490] It is clear enough, in this context, what "body-based suffering" means. But there is also an important distinction to be made between body-based suffering and mere bodily pain. See Hanna, *Kantian Ethics and Human Existence*, ch. 4.

[491] For the other three arguments, see section 4.5 above.

[492] See Hanna, *Kantian Ethics and Human Existence*, ch. 3.

[493] Indeed, it is Griffin's commitment to classical *microscopic* Whiteheadianism, and to the idea that macroscopic, manifestly real beings, including real human persons, are ultimately "societies" of microscopic, humanly unperceivable Whiteheadian "actual entities," that is the least plausible commitment of his view and Whitehead's alike—in fact, it is simply a residual Leibniz-style monadological thesis in *noumenal metaphysics*. See Griffin, *Unsnarling the World-Knot*; Rosenberg, *A Place for Consciousness*; and A.N. Whitehead, *Process and Reality* (London: Macmillan, 1929).

[494] —Which means, basically: being middle class or above, and a US citizen or permanent resident, with adequate income to pay for the kind of good health care that, in fact, *everyone*, no matter what their class or income, immigrant status, race, etc., etc., *should be receiving for free*, according to sufficient respect for human dignity. But as of September 2017, perhaps the proposed Medicare For All act provides some ground for hope.

[495] See Hanna, *Kantian Ethics and Human Existence*, ch. 3.

[496] See, for example, Olson, *The Human Animal*, for the view that "person" is a phase sortal term.

[497] See, for example, M. Spicuzza, "Doctors Separate Twin Boys Despite Finding Brains Were Fused," *New York Times* (6 August 2004), available online at URL = <http://www.nytimes.com/2004/08/06/nyregion/doctors-separate-twin-boys-despite-finding-brains-were-fused.html>.

[498] For a fascinating account of another set of janus conjoined twins, born in Canada in 2006, Krista and Tatiana Hogan, see S. Dominus, "Could Conjoined Twins Share a Mind?," *New York Times* (25 May 2011), available online at URL = <http://www.nytimes.com/2011/05/29/magazine/could-conjoined-twins-share-a-mind.html>.

[499] The crucial difference between (1) totipotent human cells between 14 and 18 days after conception or fertilization, which are *not* human individuals, and (2) the Filipino janus twins, which *do* jointly constitute a single human individual, is that the biological structure of totipotent cells *is* inherently open to fission or fusion during the natural course of its development between 14 and 18 days after conception or fertilization, whereas the organism

constituted by the Filipino janus twins is *not* inherently open to fission or fusion during the natural course of its development, even though a surgical intervention can produce an artificial fission. Many thanks to Kelly Vincent for pressing me on this point.

[500] Correspondingly, I also hold that the Canadian janus twins Krista and Tatiana Hogan jointly constitute a single real person. And, in turn, judging by the news reports and interviews, this seems to be the basic moral reason why Krista/Tatiana's parents decided against surgical separation in 2011. As of Fall 2017, Krista/Tatiana remains unseparated. Assuming that surgical separation never occurs, precisely how the rest of her amazing life plays out in personal, moral, and legal terms, remains to be seen.

[501] For an in-depth discussion of the morality of abortion and infanticide, and also of the Trolley Problem, see Hanna, *Kantian Ethics and Human Existence*, chs. 3 and 5.

[502] See J. Bayley, *Elegy for Iris* (London: Picador, 2001).

[503] See J. McMahan, *The Ethics of Killing* (Oxford: Oxford Univ. Press, 2002), p. 47.

[504] "Somebody ought to write a book about people sometime—they're peculiar." See D. Hammett, "Too Many Have Lived," in D. Hammett, *The Adventures of Sam Spade and Other Stories*, available online at URL = <http://www.fadedpage.com/showbook.php?pid=20120735>, p. 11.

[505] See Hanna and Maiese, *Embodied Minds in Action*, esp. chs. 1-2, 7, and 8.

[506] Parfit, *Reasons and Persons*, p. 200.

[507] Parfit, *Reasons and Persons*, pp. 229-230.

[508] See, for example, Buss and Overton (eds.), *The Contours of Agency: Essays on Themes from Harry Frankfurt*.

[509] A. Camus, "The Myth of Sisyphus," in Cahn and Markie (eds.), *Ethics: History, Theory, and Contemporary Issues*, pp. 397-405, at p. 405.

[510] See Williams, "Moral Luck," p. 28.

[511] W. Stevens, "The Man with the Blue Guitar," *Poetry* 50 (May 1937), canto I.

[512] Parfit, *Reasons and Persons*, p.231.

[513] Parfit, *Reasons and Persons*, p. 231.

[514] See Hanna, *Kantian Ethics and Human Existence*, sections 6.5, 6.9. and 6.10.

[515] Parfit, *Reasons and Persons*, p. 253.

[516] E. Olson, "Personal Identity," *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), E.N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2017/entries/identity-personal/>.

[517] Parfit, *Reasons and Persons*, pp. 254-255.

[518] In *Kant, Agnosticism, and Anarchism*, part 2, I distinguish between (i) The Realm of Ends per se, which is an ideal moral community containing all actual or possible finite or real persons, whether human or non-human, and (ii) The *Real* Realm of Ends, which is an actual-world-indexed, and (as far as we know) Earth-bound, cosmopolitan ethical community consisting of all and only real human persons, aka *humanity*.

[519] See Hanna and Maiese, *Embodied Minds in Action*, chs. 1-2.

[520] Parfit, *Reasons and Persons*, p. 255.

INDEX