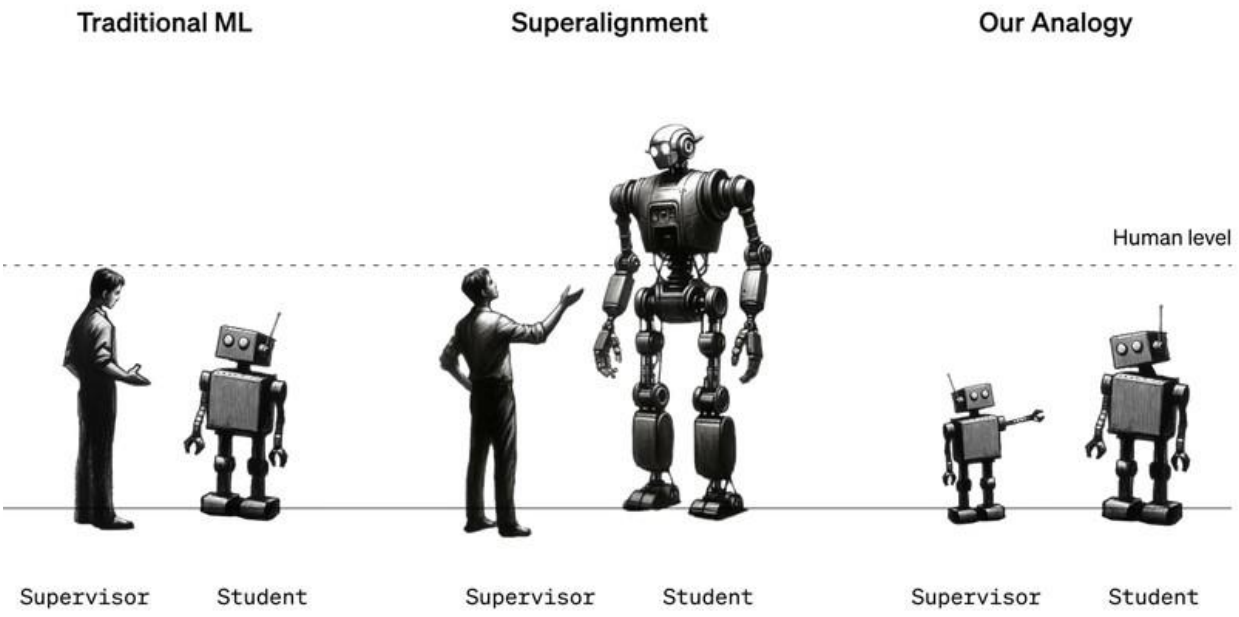
OpenAI, The Superalignment Problem, and Human Values

Robert Hanna



A simple analogy for superalignment: In traditional machine learning (ML), humans supervise AI systems weaker than themselves (left). To align superintelligence, humans will instead need to supervise AI systems smarter than them (center). We cannot directly study this problem today, but we can study a simple analogy: can small models supervise larger models (right)? (Open AI, 2023)

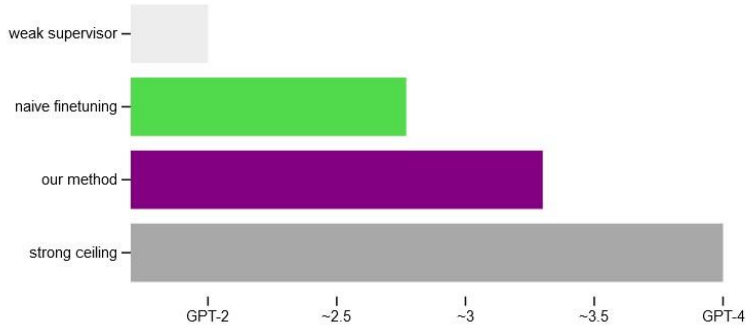
In an important paper released in mid-December 2023, an OpenAI research team—called "The Superalignment Generalization Team"—discussed what they call "the superalignment problem." Here's how OpenAI described it in an overview post on their main website:

> We believe superintelligence—AI vastly smarter than humans—could be developed within the next ten years. However, we still do not know how to reliably steer and control superhuman AI systems. Solving this problem is essential for ensuring that even the most advanced AI systems in the future remain safe and beneficial to humanity….
>
> Current alignment methods, such as reinforcement learning from human feedback (RLHF), rely on human supervision. However, future AI systems will be capable of extremely complex and creative behaviors that will make it hard for humans to reliably supervise them. For example, superhuman models may be able to write millions of lines

> of novel—and potentially dangerous—computer code that would be very hard even for expert humans to understand.
>
> Relative to superhuman AI models, humans will be "weak supervisors." This is a core challenge for AGI alignment: how can weak supervisors trust and control substantially stronger models? (OpenAI, 2023)

In other words, the superalignment problem is that Large Language Models (LLMs) like ChatGPT *will soon become so massively complex and prodigious that even the people who construct and run them won't be able to comprehend or control them.* In turn, the Superalignment Generalization team's proposed solution to the superalignment problem is what they call "weak-to-strong generalization":

> To make progress on this core challenge, we propose an analogy we can empirically study today: can we use a smaller (less capable) model to supervise a larger (more capable) model?
>
> [See the image displayed at the top of this essay. RH]
>
> Naively, we might not expect a strong model to perform better than the weak supervisor that provides its training signal—it may simply learn to imitate all the errors the weak supervisor makes. On the other hand, strong pretrained models have excellent raw capabilities—we don't need to teach them new tasks from scratch, we just need to elicit their latent knowledge. The critical question is then: will the strong model generalize according to the weak supervisor's underlying intent—leveraging its full capabilities to solve the task even on difficult problems where the weak supervisor can only provide incomplete or flawed training labels?
>
> Our results



Typical weak-to-strong generalization across NLP benchmarks: We use a GPT-2-level model as a weak supervisor to finetune GPT-4.

> We can significantly improve generalization in many settings. We use a simple method that encourages the strong model to be more confident—including confidently disagreeing with the weak supervisor if necessary. When we supervise GPT-4 with a GPT-2-level model using this method on NLP tasks, the resulting model typically performs somewhere between GPT-3 and GPT-3.5. We are able to recover much of GPT-4's capabilities with only much weaker supervision.
>
> This method is a proof of concept with important limitations; for example, it still doesn't work on ChatGPT preference data. However, we also find signs of life with other approaches, such as optimal early stopping and bootstrapping from small to intermediate to large models.
>
> Collectively, our results suggest that (1) naive human supervision—such as reinforcement learning from human feedback (RLHF)—could scale poorly to superhuman models without further work, but (2) it is feasible to substantially improve weak-to-strong generalization. (OpenAI, 2023)

More specifically and technically, however, here's the abstract for the Superalignment Generalizations team's research paper:

> Widely used alignment techniques, such as reinforcement learning from human feedback (RLHF), rely on the ability of humans to supervise model behavior—for example, to evaluate whether a model faithfully followed instructions or generated safe outputs. However, future superhuman models will behave in complex ways too difficult for humans to reliably evaluate; humans will only be able to weakly supervise superhuman models. We study an analogy to this problem: can weak model supervision elicit the full capabilities of a much stronger model? We test this using a range of pretrained language models in the GPT-4 family on natural language processing (NLP), chess, and reward modeling tasks. We find that when we naively finetune strong pretrained models on labels generated by a weak model, they consistently perform better than their weak supervisors, a phenomenon we call weak-to-strong generalization. However, we are still far from recovering the full capabilities of strong models with naive finetuning alone, suggesting that techniques like RLHF may scale poorly to superhuman models without further work. We find that simple methods can often significantly improve weak-to-strong generalization: for example, when finetuning GPT-4 with a GPT-2-level supervisor and an auxiliary confidence loss, we can recover close to GPT-3.5-level performance on NLP tasks. Our results suggest that it is feasible to make empirical progress today on a fundamental challenge of aligning superhuman models. (Burns et al., 2023: p. 1)

In other words, the Superalignment Generalization team's proposed solution to the superalignment problem is to use *less* complex and *less* prodigious LLMs in order to comprehend and control the *more* complex and *more* prodigious LLMs.

But there's an obvious philosophical objection to this methodological strategy. Let's suppose that it's *actually* the case that using the currently less complex and less prodigious LLMs that are being aligned by reinforcement learning from human feedback, is *already* rationally unjustified and immoral. Then how could using *those* LLMs ever possibly comprehend and control the massively more complex and prodigious future LLMs in a rationally and morally justified way? On the last page of the paper, the team says:

> • No need to solve human values. We assume we do not need to solve hard philosophical questions of human values and value aggregation before we can align a superhuman researcher model well enough that it avoids egregiously catastrophic outcomes. (Burns et al., 2023: p. 49)

In other words, implementing the Superalignment Generalization team's solution will be completely unconstrained by any annoying and distracting philosophical worries to the effect that what OpenAI and other LLM research programs are *currently* doing is in fact *already* rationally unjustified and immoral—indeed, as rationally unjustified and immoral as the US government's Manhattan Project during World War II for building, testing, and using the atomic bomb.

Moreover, at least some people, including Geoffrey Hinton and myself, have expressed precisely these philosophical worries:

> On 1 May 2023—significantly, *May Day*—Geoffrey Hinton, a groundbreaking researcher on neural networks and LLMs, quit his job at Google and then publicly made [the following point,] explicitly using the analogy with The Manhattan Project:
>
> > As companies improve their A.I. systems, [Hinton] believes, they become increasingly dangerous. "Look at how it was five years ago and how it is now," he said of A.I. technology. "Take the difference and propagate it forwards. That's scary."
> >
> > Until last year, he said, Google acted as a "proper steward" for the technology, careful not to release something that might cause harm. But now that Microsoft has augmented its Bing search engine with a chatbot — challenging Google's core business — Google is racing to deploy the same kind of technology. The tech giants are locked in a competition that might be impossible to stop, Dr. Hinton said.
> >
> > His immediate concern is that the internet will be flooded with false photos, videos and text, and the average person will "not be able to know what is true anymore."
> >
> > He is also worried that A.I. technologies will in time upend the job market. Today, chatbots like ChatGPT tend to complement human workers, but they could replace paralegals,

> personal assistants, translators and others who handle rote tasks. "It takes away the drudge work," he said. "It might take away more than that."
>
> Down the road, he is worried that future versions of the technology pose a threat to humanity because they often learn unexpected behavior from the vast amounts of data they analyze. This becomes an issue, he said, as individuals and companies allow A.I. systems not only to generate their own computer code but actually run that code on their own. And he fears a day when truly autonomous weapons—those killer robots—become reality.
>
> "The idea that this stuff could actually get smarter than people—a few people believed that," he said. "But most people thought it was way off. And I thought it was way off. I thought it was 30 to 50 years or even longer away. Obviously, I no longer think that."
>
> Many other experts, including many of his students and colleagues, say this threat is hypothetical. But Dr. Hinton believes that the race between Google and Microsoft and others will escalate into a global race that will not stop without some sort of global regulation.
>
> But that may be impossible, he said. Unlike with nuclear weapons, he said, there is no way of knowing whether companies or countries are working on the technology in secret. The best hope is for the world's leading scientists to collaborate on ways of controlling the technology. "I don't think they should scale this up more until they have understood whether they can control it," he said.
>
> Dr. Hinton said that when people used to ask him how he could work on technology that was potentially dangerous, he would paraphrase Robert Oppenheimer, who led the U.S. effort to build the atomic bomb: "When you see something that is technically sweet, you go ahead and do it."
>
> He does not say that anymore. (New York Times, 2023)

What Hinton (apparently) and I (absolutely) strongly believe, then, is that *we ought to ban all giant AI experiments and LLM/chatbot technology while they are still in their infancy*, just as *we ought to have banned all atomic bomb experiments and nuclear weapons technology while they were still in their infancy.* (Hanna, 2023i: pp. 5-6)

The reasoning justifying this strong belief flows from the doctrine I call "dignitarian neo-Luddism with respect to digital technology":

> [B]y *digital technology* I mean all mechanical technology that inherently involves computers, algorithms, digital data or information, artificial intelligence/AI, or robotics. Then, *neo-Luddism with respect to digital technology* says that
>
> > not all digital technology is bad and wrong, but instead all and only the digital technology that harms and oppresses ordinary people (i.e., people other than

> digital technocrats) is bad and wrong, and therefore all and only this bad and wrong digital technology should be rejected but not—except in extreme cases of digital technology whose coercive use is actually violently harming and oppressing ordinary people, for example, digitally-driven weapons or weapons-systems being used for mass destruction or mass murder—destroyed, rather only either simply refused, non-violently dismantled, or radically transformed into its moral opposite.

Finally, *dignitarian neo-Luddism with respect to digital technology* says that

> not all digital technology is bad and wrong, but instead all and only the digital technology that harms and oppresses ordinary people (i.e., people other than digital technocrats), by either failing to respect our human dignity sufficiently or by outright violating our human dignity, is bad and wrong, and therefore all and only this bad and wrong digital technology should be rejected but not—except in extreme cases of digital technology whose coercive use is actually violently harming and oppressing ordinary people, for example, digitally-driven weapons or weapons-systems being used for mass destruction or mass murder—destroyed, rather only either simply refused, non-violently dismantled, or radically transformed into its moral opposite.

It needs to emphasized and re-emphasized that just as I'm *pro-science* while also being *anti-frankenscience*, so too dignitarian neo-Luddism with respect to digital technology is *also* committed to the *positive* dignitarian moral doctrine that *some* digital technology is permissible and good, and therefore *ought to be used*, precisely because it's morally unobjectionable and promotes the betterment of humankind, and more generally sufficiently respects human dignity. For example, in my opinion this is true of posting or self-publishing essays about dignitarian digital/AI ethics for universal free sharing on the internet. Why else would I be doing it? But in the context of this essay, I'm focusing primarily on the negative dignitarian moral doctrine.

So much for the definitions, and now for some moral imperatives. What I strongly believe is that *we all ought to be dignitarian neo-Luddites with respect to digital technology*. Why? To be sure, there are many ways in which digital technology can be bad and wrong in the dignitarian sense, including invasive digital surveillance, digitally-driven weapons and weapon systems, algorithmic bias, and digital manipulation and nudging. And of course there are also ways in which digital technology can be bad and wrong in the utilitarian sense, for example, putting many people out of work. But the principal reason for being a dignitarian neo-Luddite with respect to digital technology is that *our excessive use of and indeed addiction to digital technology is systematically undermining our innate capacities for thinking, caring, and acting for ourselves*. This is *preeminently* true with respect to the new chatbots—for example, ChatGPT and LaMDA—and what I've called *the myth of artificial intelligence* more generally (see, e.g., Hanna, 2023a, 2023b, 2023c, 2023d, 2023e, 2023f), but

also *to an increasingly important degree* true for our excessive use of and addiction to smart-phones, desktop and laptop computers, the internet, social media, and so-on and so-forth. When you combine our excessive use of and addiction to chatbots and AI with our excessive use of and addiction to smart-phones, desktop and laptop computers, the internet, social media, etc., the result is nothing less than *an all-out existential attack on our rational human mindedness or intelligence.*

By "our rational human mindedness or intelligence" I mean the essentially embodied, unified set of basic innate cognitive, affective, and practical capacities present in all and only those human animals possessing the essentially embodied neurobiological basis of those capacities, namely: (i) *consciousness*, i.e., subjective experience, (ii) *self-consciousness*, i.e., consciousness of one's own consciousness, second-order consciousness, (iii) *caring*, i.e., desiring, emoting, or feeling, (iv) *sensible cognition*, i.e., sense-perception, memory, or imagination, (v) *intellectual cognition*, i.e., conceptualizing, believing, judging, or inferring, (vi) *volition*, i.e., deciding, choosing, or willing, and (vii) *free agency*, i.e., free will and practical agency. This unified set of capacities constitutes our *human real personhood*, which in turn is *the metaphysical ground of our human dignity* (Hanna, 2023g, 2023h). Therefore, this all-out existential attack on our rational human mindedness or intelligence is also an all-out existential attack on our human dignity.

The Cassandra-like prophecy and warning that I'm issuing, however, is *not* that chatbots or AI more generally could ever become rational, super-intelligent, and morally satanic, and then run amok. Interestingly, that seems to be one of Hinton's worries. But in fact, it's *metaphysically impossible* for computing machines ever to be rationally minded or intelligent in the sense that *we're* rationally minded or intelligent, because (i) it's *metaphysically necessary* that all creatures possessing the seven basic innate capacities I listed above are complex living organisms, i.e., *animals* (Hanna and Maiese, 2009), hence not *machines*, hence not *computing machines*, and (ii) it's also *metaphysically necessary* that our rational mindedness or intelligence includes (iia) an innate non-basic ("non-basic," in the sense that it essentially depends on the seven basic innate capacities listed in the just-previous paragraph) capacity for *spontaneous creativity*, and also (iib) an innate non-basic capacity for either conceptual or essentially non-conceptual *a priori intuition* of (iib1) innately-specified universal, unconditional, a priori or non-empirical moral principles such as *everyone ought always to choose and act with sufficient respect for everyone's dignity, including their own*, (iib2) universal, unconditional, a priori or non-empirical logical principles such as the minimal principle of non-contradiction, namely, *not every statement is both true and false*, and (iib3) the universal a priori or non-empirical formal structures of the orientable, three-dimensional space and the forward-directed, processual, purposive, asymmetric organic time in which our minded animal bodies are ineluctably embedded (Hanna, 2006, 2015, 2018), *none of which can ever exist in computing machinery*. And, closely related to these metaphysically necessary modal facts, there are also various other linguistic, logical, mathematical, and metaphysical reasons why computing machinery can never be rationally minded or intelligent in the sense that *we're* rationally minded or

intelligent (see, e.g., Chomsky, Roberts, and Watamull, 2023; Hanna, 2023a, 2023b, 2023c, 2023d, 2023e, 2023f; Keller, 2023; Landgrebe and Smith, 2022).

On the contrary, the Cassandra-like prophecy and warning that I'm issuing about this all-out existential attack on our rational human mindedness or intelligence is instead directed at *the global technocratic capitalist corporations—especially those that supply weapons and surveillance systems for military and government use—millionaires, and billionaires* who reap immense profits and wield immense political power by designing, producing, marketing, and above all *controlling* our use of and reliance on digital technology: namely, the members of the military-industrial-digital complex. Correspondingly, my Cassandra-like prophecy and warning is simply this:

> the members of the military-industrial-digital complex are systematically harming and oppressing ordinary people like us by not only enabling but also effectively mandating our excessive use of and addiction to digital technology, which in turn systematically undermines our innate capacities for thinking, caring, and acting for ourselves, and therefore undermines our human real personhood, and thereby violates our human dignity—therefore, we ought to ban all giant AI experiments and LLM/chatbot technology while they are still in their infancy, just as we ought to have banned all atomic bomb experiments and nuclear weapons technology while they were still in their infancy. (Hanna, 2023i: pp. 7-10)

For all those reasons, I strongly believed those critical claims when I wrote them in July 2023, and I even more strongly believe them now. Correspondingly, by assuming without argument that "we do not need to solve hard philosophical questions of human values and value aggregation before we can align a superhuman researcher model well enough that it avoids egregiously catastrophic outcomes" (Burns et al., 2023: p. 49), the Superalignment Generalization team's appeal to the method of weak-to-strong generalization *does absolutely nothing* either to address and or to respond adequately and effectively to my worries.

By way of elaborating that point, I want to explore the analogy between OpenAI's proposed solution to the superalignment problem and the Manhattan Project just a little further. No doubt, in July 1945, Robert Oppenheimer made the assumption that "we do not need to solve hard philosophical questions of human values and value aggregation before" the Manhattan Project could scientifically enable President Harry Truman to drop the atomic bomb on hundreds of thousands of Japanese civilians in Hiroshima and Nagasaki in August 1945, in order to avoid "egregiously catastrophic outcomes" for the USA during a land invasion of Japan. But arguably, dropping the atomic bomb on two Japanese cities was nevertheless a crime against humanity, that went unprosecuted only because the USA won the brutal Pacific War, and exacted "victor's justice" by staging the

Tokyo war crimes trials (see, e.g., Bass, 2023). So too, arguably, the Superalignment Generalization team's proposing the method of weak-to-strong generalization as a solution to the superalignment problem, so that OpenAI and other global technocratic capitalist corporations can unconstrainedly continue to develop and market LLMs in order to reap immense profits and wield immense political power, is another crime against humanity.

Now, to conclude: Obviously, the atomic bomb wasn't and isn't superintelligent. Nevertheless, by virtue of its massive and uncontrollable destructive power, it was and is an existential threat to humankind *that could have been avoided*. So too, even if, as I've argued, current and future LLMs *are not* and *cannot ever be* superintelligent, precisely because they are not and cannot ever be intelligent in the sense in which we're intelligent (Chomsky, Roberts, and Watamull, 2023; Hanna, 2023a, 2023b, 2023c, 2023d, 2023e, 2023f; Keller, 2023; Landgrebe and Smith, 2022), nevertheless, by virtue of our excessive use of and addiction to digital technology, together with massively more complex and uncontrollable future LLMs looming on the temporal horizon like a towering mushroom cloud, they're another existential threat to humankind *that still can be avoided*.

REFERENCES

(Bass, 2023). Bass, G. *Judgment at Tokyo: World War II on Trial and the Making of Modern Asia.* New York: Alfred A. Knopf.

(Burns et al., 2023). Burns, C., Izmailov, P. Kirchner, J.H., Baker, B. Gao, L., Aschenbrenner, L. Chen, Y. Ecoffet, A., Joglekar, M. Sutskever. J.L.I., and Wu, J. "Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision." *OpenAI*. 14 December. Available online at URL = <https://cdn.openai.com/papers/weak-to-strong-generalization.pdf>.

(Chomsky, Roberts, and Watumull, 2023). Chomsky, N., Roberts, I. and Watumull, J. "Noam Chomsky: The False Promise of ChatGPT." *New York Times*. 8 March. Available online at URL = <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>.

(Hanna, 2006). Hanna, R. *Rationality and Logic*. Cambridge MA: MIT Press. Also available online in preview at URL = <https://www.academia.edu/21202624/Rationality_and_Logic>.

(Hanna, 2015). Hanna, R. *Cognition, Content, and the A Priori: A Study in the Philosophy of Mind and Knowledge.* THE RATIONAL HUMAN CONDITION, Vol. 5. Oxford: Oxford Univ. Press. Also available online in preview at URL = <https://www.academia.edu/35801833/The_Rational_Human_Condition_5_Cognition_Content_and_the_A_Priori_A_Study_in_the_Philosophy_of_Mind_and_Knowledge_OUP_2015_>.

(Hanna, 2018). Hanna, R. *Deep Freedom and Real Persons: A Study in Metaphysics.* THE RATIONAL HUMAN CONDITION, Vol. 2. New York: Nova Science. Available online in preview at URL = <https://www.academia.edu/35801857/The_Rational_Human_Condition_2_Deep_Freedom_and_Real_Persons_A_Study_in_Metaphysics_Nova_Science_2018_>.

(Hanna, 2023a). Hanna, R. "How and Why ChatGPT Failed The Turing Test." Unpublished MS. Available online at URL = <https://www.academia.edu/94870578/How_and_Why_ChatGPT_Failed_The_Turing_Test_January_2023_version_>.

(Hanna, 2023b). Hanna, R. "It's All Done With Mirrors: A New Argument That Strong AI is Impossible." Unpublished MS. Available online at URL = <https://www.academia.edu/95296914/Its_All_Done_With_Mirrors_A_New_Argument_That_Strong_AI_is_Impossible_January_2023_version_>.

(Hanna, 2023c). Hanna, R. "Are There Some Legible Texts That Even The World's Most Sophisticated Robot Can't Read?" Unpublished MS. Available online at URL = <https://www.academia.edu/95866304/Are_There_Some_Legible_Texts_That_Even_The_Worlds_Most_Sophisticated_Robot_Cant_Read_January_2023_version_>.

(Hanna, 2023d). Hanna, R. "Babbage-In, Babbage-Out: On Babbage's Principle." Unpublished MS. Available online at URL = <https://www.academia.edu/101462742/Babbage_In_Babbage_Out_On_Babbages_Principle_May_2023_version_>.

(Hanna, 2023e). Hanna, R. "The Myth of Artificial Intelligence and Why It Persists." Unpublished MS. Available online at URL = <https://www.academia.edu/101882789/The_Myth_of_Artificial_Intelligence_and_Why_It_Persists_May_2023_version_>.

(Hanna, 2023f). Hanna, R. "Creative Rage Against the Computing Machine: Necessary and Sufficient Conditions for Authentic Human Creativity." Unpublished MS. Available online at URL = <https://www.academia.edu/101633470/Creative_Rage_Against_the_Computing_Machine_Necessary_and_Sufficient_Conditions_for_Authentic_Human_Creativity_June_2023_version_>.

(Hanna, 2023g). Hanna, R. "Dignity, Not Identity." Unpublished MS. Available online at URL = <https://www.academia.edu/96684801/Dignity_Not_Identity_February_2023_version_>.

(Hanna, 2023h). Hanna, R. "In Defence of Dignity." *Borderless Philosophy* 6: 77-98. Available online at URL = <https://www.cckp.space/single-post/bp6-2023-robert-hanna-in-defence-of-dignity-77-98>.

(Hanna, 2023i). Hanna, R. "Oppenheimer, Kaczynski, Shelley, Hinton, & Me: Don't Pause Giant AI Experiments, *Ban Them*." Unpublished MS. Available online at URL = <https://www.academia.edu/97882365/Oppenheimer_Kaczynski_Shelley_Hinton_and_Me_Don_t_Pause_Giant_AI_Experiments_Ban_Them_July_2023_version_>.

(Hanna and Maiese, 2009). Hanna, R. and Maiese, M., *Embodied Minds in Action*. Oxford: Oxford Univ. Press. Available online in preview at URL = <https://www.academia.edu/21620839/Embodied_Minds_in_Action>.

(Keller, 2023). Keller, A. "Artificial, But Not Intelligent: A Critical Analysis of AI and AGI." *Against Professional Philosophy*. 5 March. Available online at URL = <https://againstprofphil.org/2023/03/05/artificial-but-not-intelligent-a-critical-analysis-of-ai-and-agi/>.

(Landgrebe and Smith, 2022). Landgrebe, J. and Smith, B. *Why Machines Will Never Rule the World: Artificial Intelligence without Fear*. London: Routledge.

(New York Times, 2023). Metz, C. "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead." *The New York Times*. 1 May. Available online at URL = <https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html>.

(OpenAI, 2023). *OpenAI*. "Research: Weak-to-Strong Generalization." Available online at URL = <https://openai.com/research/weak-to-strong-generalization>.