This is a draft, pre-peer review copy and not the finalized manuscript. Please cite the final version, which will appear in the book, *Computers with Salaries and Cemeteries: AI Ethics from Industry to Philosophy to Science Fiction* (forthcoming)

What Good is Superintelligent AI?

Abstract:

Extraordinary claims about both the imminence of superintelligent AI systems and their foreseen capabilities have gone mainstream. It is even argued that we should exacerbate known risks such as climate change in the short term in the attempt to develop superintelligence (SI), which will then purportedly solve those very problems. Here, I examine the plausibility of these claims. I first ask what SI is taken to be and then ask whether such SI could possibly hold the benefits often envisioned. I conclude that we do not have sufficient reason to believe that we are close enough to developing SI capable of resolving major human problems to justify taking on substantial risks in the attempt to develop it.

Keywords:

Superintelligence; AI; Value of superintelligence; AGI; human intelligence; animal intelligence

1. Introduction

Extraordinary claims relating to superintelligence (SI) abound. Enthusiasts not only believe SI to be imminent but also have exceptionally strong faith in its value. It is claimed that SI will do things like "fixing the climate, establishing a space colony, and [discovering] of all of physics", and bring about "massive prosperity" (Altman, 2024). Some even propose the risky strategy of accelerating current AI development — with its massive energy and water needs — in the hope that this will result in SI, with the promise that it will then be able to resolve the climate crisis (Niemeyer, 2024). Here, I

explore whether we have good reason to believe such claims. Do we have good reason to believe that SI is imminent and that it will be of such unimaginable benefit to all of humanity, to the point that we should take on massive risks to achieve it? I order to address these questions, we first need clarity on what "superintelligence" is supposed to be, which is a fraught question in itself. To get to grips with it, it may help to examine what "intelligence" is taken to be. To do this, we can start with a kind of intelligence that we are familiar with: human intelligence.

2. Human intelligence

There is no generally-agreed-upon definition of human intelligence. Luckily, for present purposes, describing the kinds of high-level competences that humans have will suffice. These include: the ability to engage in complex goal-seeking behavior, in complex communication, and in complex cooperation, amongst others (see Schwitzgebel & Pober, 2024). Such capabilities certainly seem to require intelligence. To successfully engage in them requires, *inter alia*, the ability to understand the (human) world, as well as human language, behaviour, and goals. They also require the ability to formulate or acquire goals, the inclination to act in ways that would allow for the achievement of these goals, and the capacity to understand what kinds of behaviours would be required, given the state of the world and the other humans involved, and the ability to plan and act accordingly.[1] Clearly, if we wanted to create artificial intelligence (AI) systems with (only) human-level intelligence (i.e. artificial general intelligence (AGI)), it would at least need to be able to do all of these things as well as humans generally tend to. If we wanted to create SI, it would need to have superior capabilities (compared to humans) in many of these areas.

As the above sketch shows, even achieving only human-level AGI will be quite a complex endeavour. Whereas current AI systems tend to be able to match, and sometimes exceed, human capabilities in specific tasks (e.g. generating plausible-

[1] I leave it open whether these capacities require some form of consciousness.

sounding text, playing chess, or predicting protein structures), none of them can approximate human-level capabilities in all human tasks. In all probability, getting from these systems to AGI will require remarkable feats of engineering, and it would be extremely expensive. Nevertheless, such systems seem possible, in principle, and there could be great value in creating them. AGI that matches general human performance in most tasks could reliably be deployed to perform those tasks, with concomitant benefits, including undertaking dangerous tasks and drudge work or, under some conceptions of value, replace human workers cost-effectively.[2] Presumptively, at least, there seem to be good reasons to attempt to develop human-like AGI and even to take on some (not all) of the risks that this may entail. Does this line of reasoning carry over to SI?

3. Superintelligence

Much of the discourse on both AI and SI lacks the conceptual clarity and rigour that one would like from scientifically-based discourse (Mitchell, 2024). Too often, this results in unhelpful handwaving and intuition-based prognoses, which are not good bases for informed decision making. Unsurprisingly, it is not always clear what is meant with "superintelligence"; moreover, the term takes on different connotations in different contexts. Bostrom (2006, p. 11) gives a canonical description of the kind of thing "superintelligence" is often taken to be:

> By a "superintelligence" we mean an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills. This definition leaves open how the superintelligence is implemented: it could be a digital computer, an ensemble of networked computers, cultured cortical tissue or what have you. It also leaves open whether the superintelligence is conscious and has subjective experiences.

[2] All of this is predicated on the assumption that we can create human-like AGI without having to replicate human consciousness and the related physiological and psychological characteristics that enable this in humans. Here, I am taking an agnostic stance on the issue.

As the term is currently used, it often denotes the above characteristics, but with the added assumption that it will be implemented in large language model (LLM)-based or similar technologies, which, as we shall see, we have reason to be skeptical about.

Two further, related questions that also merit skeptical attention is the plausibility of i) claims relating to the capacities future SI is predicted to have, and ii) claims that such SI will "solve" all manner of seemingly intractable human problems. The answers to these two questions would go some way towards helping us determine whether we should seriously contemplate taking major risks in the short term, such as exacerbating climate change, in the quest to develop SI.

3.1 What exactly is SI?

Clearly, SI is meant to entail some form of exceptional intelligence which is much greater than any intelligence that we are familiar with. But what could this mean? One underlying assumption seems to be that intelligence, whatever it is, exists on a continuum, with SI being at a (desirable) higher capability level on that continuum than human-level intelligence. However, this view is far from obvious. There are many extremely diverse ways to be intelligent. This is best illustrated by considering non-human animals. Many of us are quite willing to attribute intelligence to some or other subset of animals, including other primates, mammals, birds, reptiles, fish, or perhaps even insects. Indeed, for a common conception of machine intelligence — *having the ability to entertain goals and act in pursuit of them* (Legg & Hutter, 2007) — all of these entities are intelligent, as are amoeba, macromolecules, and even thermostats (see Dennett, 2000). On a more refined definition of intelligence by Andrews (2010) — *exhibiting capacities for flexible, goal-oriented behaviour through information processing* — we retain animals and perhaps, depending on our conception of "flexible", all living organisms, and some forms of AI.

When considering these examples, it becomes clear that thinking of intelligence as falling along a single continuum where more overall intelligence is better than less is a vast oversimplification. Animals exhibit a range of cognitive capacities that are not available to humans and that serve them well in their own contexts. Clearly, not all cognitive capacities are present in all animals, nor are all shared capacities present to the same degree in different species. Not all cognitive capacities would be equally useful to all animals. Cognitively and behaviourly speaking, caterpillars, bats, and chimpanzees show many striking differences; yet, their different capacities tend to serve them well enough to obtain their goals (Tye, 2016).[3] Hence, animals exhibit cognitive capacities (instantiations of intelligence) that might better be thought of as falling along many different dimensions. It is plausible to suggest that being more capable among a dimension relevant to a particular individual's context could be beneficial to it, but both the specific capacities and the particular context matter. It is not obvious that just having more overall intelligence (whatever that might be) would inevitably be beneficial. It may even be detrimental, if the additional cognitive capacities require more energy or lead to expanded access to information, which might take longer to sift through and assess (see Dennett, 2024).

Human intelligence shares in the same high-level evolutionary constraints as animal intelligence, but we also occupy a unique cognitive niche, presumably due to language (Dennett, 2000; Spelke, 2022). It is likely, as Kant (1908) already argued in 1781 that some aspects of human cognition are innate and precede language acquisition (Lake et al., 2017; Spelke, 2000). Human language then seems to allow us to build on this core knowledge, which accounts for some of the ways in which human intelligence differs from that of animals. Thus, humans and animals all occupy a range of positions in the space of possible ways to be intelligent, and there may be countless others. It seems very plausible to suggest that not all of these possible cognitive

[3] In evolved creatures, intelligence tends to be useful in as far as it enables its holder to meet their evolutionary-endowed goals (roughly, obtaining energy and other resources, avoiding danger, and successfully procreating).

capacities would be useful in all contexts. The multi-dimensionality of intelligence also makes apparent why it is so difficult to assess intelligence in other kinds of entities. Bee intelligence, starfish intelligence, and mole intelligence can, in many respects, seem quite alien from ours (Cartmill, 2023; Dennett, 1995). As Dennett (2000) points out, even domesticated pets, whom we can have great affinity with, sometimes show baffling (to us) gaps in their intelligence. Nevertheless, non-human animals share at least some underlying mechanistic and structural similarities to humans, which allows us to make relatively informed conjectures about their goals and behaviours. This need not be the case with AI.[4]

The first implication of the above outline of the multi-dimensionality of intelligence is that it matters what *kind* of intelligence we have in mind when talking about SI. In what respect do we expect our SI to be superlatively intelligent? The second implication is that context matters. If we are not simply interested in SI for SI's sake, if we want to harness the capacities of SI for quite specific tasks, we have to try and determine whether the kind of superlative intelligence that we could reasonably foresee developing will be up to these tasks.

To be sure, it is *logically* possible that a SI could exist that has superlative capacities along all possible dimensions of intelligence to some or other upper bound. This is what many SI enthusiasts seem to have in mind: an ultra-intelligent entity that has maximal total possible intelligence. Possibly, this next-to-infinitely-capable SI would be able to solve any soluble problem. Conceivably, it would be able to sift through unimaginably vast amounts of information and run a gargantuan number of computations, allowing it to eventually hit upon breakthrough insights that elude humans. Still, such an SI would not be much use for solving our major challenges in the foreseeable future.

[4] As Bostrom (2012) points out, artificial SI could be far more alien to us than any possibly existing space aliens, provided that the latter represent some kind of biological creature that has arisen through evolutionary processes and are subject to broadly the same kinds of constraints as earthly creatures.

There are at least three reasons for claiming that maximally intelligent SI will not help us much. For one thing, we have no idea how to go about building it. Current AI systems excel at a limited range of specific, human-like cognitive capacities, but it remains an open question whether we can bootstrap even human-level AGI from current approaches (Mahowald et al., 2024). Current AIs exhibit many surprising cognitive deficits (Krakovna et al., 2020; Lehman et al., 2020), to the extent that many theorists are hesitant to attribute intelligence to them (Bender & Koller, 2020; Floridi, 2023; Lenat & Marcus, 2023). Secondly, it is even more unclear how we are to get from our current systems to a maximally intelligent SI. Thirdly, such a system would require enormous amounts of computing power as well as vast amounts of time to run its calculations. The solutions it eventually comes up with may very well be too late to resolve a challenge like climate change. For a SI to be of use on human timescales, we would require a much more limited system which, like all intelligent creatures we know of, is able to differentiate between relevant and irrelevant information for a particular problem to efficiently deal with that information (see Dennett, 2022).

It seems that for SI to possibly be useful to us in the foreseeable future, it needs to be more limited than a maximally intelligent system. We need it to be able to act in quite a particular context: the human world, with its human problems, on human timescales, and within given resource constraints. Hence, our SI needs to be constrained along sufficiently human dimensions of intelligence, while also being better than humans at the kinds of capabilities relevant to our challenges. But which capabilities would those be, and what would it mean to be "better" than humans?[5] As we have seen, Bostrom (2014) foresees SI that is *better* at "scientific creativity, general wisdom and social skills" than humans. Intuitively, these capacities seem useful for addressing problems like poverty and global warming. It also seems plausible to suggest that being "better" at these things can be useful. Nevertheless, one is hard-pressed to find specifics that allow us to properly evaluate these intuitions. Generally, there is very little by way of

[5] Cf. the 2018 science fiction series "Better than Us" which explores this topic in the context of a near future where robots abound.

explanation of just how such superhuman capacities will ensure that our most significant challenges will be resolved.

3.2 What could SI do?

The issues raised above have some overlap with questions that often receive greater attention in the SI literature, but the focus here is on how SI might harm us rather than benefit us. The *alignment problem* is generally framed as follows: SI will so far outstrip human intelligence that it will be incomprehensible to us and uncontrollable by us. This will pose an "existential" risk, where SI could bring about human extinction, by either 1) seeing us as a threat or nuisance and eliminating us, or 2) through the sheer efficacy with which it might meet its goals, accompanied by a lack of understanding of human normative and common-sensical constraints (e.g. Bostrom, 2014).[6] It is nevertheless taken for granted that SI could ultimately be so beneficial that it is worth such risks, provided that we work to *align* it. Alignment then consists in taking appropriate steps to ensure a SI is able to understand human intentions and remains benevolent towards us.[7] Such discussions usually presuppose a maximally intelligent SI like the one discussed above, which is also why the alignment problem is often dismissed out of hand. It is argued that a maximally intelligent system would inevitably also understand human intentions and goals and thus cannot be misaligned (see Müller & Cannon, 2021). This argument has some merit but does not concern us here, since my argument is that this kind of SI is not only unlikely in the foreseeable future but would, in any event, not be what is needed to address our urgent problems. Any actually possible,

[6] This is the gist of Bostrom's (2014) paperclip maximiser thought experiment. Here, the SI is faithfully fulfilling the goal its creators have given it — maximize paperclips! — which it does by using all the matter in the universe. Absent explicitly being given an additional subgoal of not harming humanity in the process, there is no reason — from its perspective — for not doing so.

[7] Posed in this way, several significant problems arise. Vold and Harris (2023) mention a few: 1) the question of which "human values" SI should be aligned with, 2) the difficulty, from a programming perspective, of incorporating a likely plurality of relevant values, 3) potential conflict between the values, giving rise to moral dilemmas, 4) identifying and conveying our values precisely enough to avoid unintended or perverse outcomes, 5) the imperfection of our current values, and 6) determining whether or not an AI system is aligned.

potentially useful SI would need to be of a lesser kind, but this in itself gives rise to a watered-down version of the alignment problem, as I discuss below.

The question remains, do we have good reason to presuppose that an SI that could be of value to us is possible within a reasonable timeframe? As already mentioned, current AI systems exhibit surprising cognitive deficits. There is extensive literature on the limitations of deep learning and of LLM-based architectures (e.g. Marcus & Davis, 2019; Bender & Koller, 2020; LeCun & Courant, 2022; Lenat & Marcus, 2023). Dung (2023), for example, provides a fascinating overview of misalignment in current AI systems where systems do not perform as anticipated or where system outcomes have unforeseen negative consequences. He also discusses the complexities involved in trying to correct for mismatches between the intentions of designers and the behaviours of systems such as LLMs and game-playing agents and comes to the sobering conclusion that misalignment might be the *default* outcome of the current deep-learning paradigm. Many of these unforeseen outcomes can be put down to the significant differences between humans and AI systems in terms of their underlying cognitive architectures, training methods, modes of learning, and the like. There is general agreement among experts that we do not yet have AGI, and there is general disagreement both about whether and when we might expect AGI and on how we might implement it (Müller & Bostrom, 2016). Given the discussion above, it should be clear that advanced capabilities in current systems — to the point where they outperform humans in certain tasks — do not necessarily entail that they can obtain capabilities along other relevant dimensions of human intelligence. In the absence of a plausible grounding theory, we do not have good reason to assume capacity for fluent language use, for example, can translate into capacity for social skills or scientific creativity or any other relevant human capabilities.

One way in which we might ensure that future AGI is human-like enough for us to be sure that it will not fail in very unhuman-like ways, and to ensure that it has the capacity

to access and understand the human world, would be to emulate the human brain.[8] The human brain is enormously complex, however, and human cognition remains poorly understood (Mandelbaum, 2022). In the absence of a comprehensive theory of how the brain works and of how many human capacities come about, the best way to ensure that we emulate the brain might be to build a replica. An artificial brain that is a close enough fit should be functionally equivalent to a biological one. But what would constitute a "close enough" fit? If an artificial brain could only be functionally equivalent to a human one if it were to be an exact copy of the human brain, the technical complexity of such a project far exceeds our current understanding and capabilities. Even if a less complex system would sufficiently emulate human cognition, it would likely necessitate an engineering feat that is beyond our current understanding or capabilities. Either way, we would need to radically extend our timeline for reaching AGI, and the question of how we get to SI would still remain.

There are suggestions other than brain emulation for reaching AGI and ultimately SI, which include direct programming, machine learning, and artificial evolution (Chalmers, 2010). All of these face similar challenges to our emulation example — questions about technical feasibility and how we will then get from AGI to SI. A guiding tenet prevalent in the SI discussion is that once human-level AGI is reached, a positive feedback loop will develop, allowing humans and AI and eventually AI itself to develop ever more capable AI. This would then lead to an "intelligence explosion" / "singularity" and eventually result in SI (Good, 1966; Chalmers, 2010; Bostrom, 2014). Yet, skeptics question this hypothesis and argue against various assumptions behind it (see Thorstad, 2024). Nevertheless, even if SI were to result from one of these approaches, questions about its capacities would remain. We have seen that in the foreseeable future, maximally intelligent SI is unlikely and may be of dubious value. Worryingly, any type of more constrained SI, while more likely, could lack crucial capacities which may diminish its ability to address our critical challenges.

[8] It is likely that even this is a massive oversimplification and that we might need to simulate an entire human body (see Dennett, 2000).

Davis (2015) correctly points out the messianic quality in some of the discussions on SI, where great intelligence is taken to entail "virtual omnipotence" or, at least, quasi-omnipotence. This is evident in claims about an SI-mediated future like the following: "[h]uman aging and illness will be reversed; pollution will be stopped; world hunger and poverty will be solved" (Kurzweil, 2005). Yet, how plausible is it to think that SI will be able to do these things? The seeming intractability of these types of problems is partly due to their nature of entailing massive coordination, where many entities who habitually act in their own (short-term) self-interest need to be persuaded to act differently, in ways that (we hope) will be to the benefit of all. The faith that SI will be able to resolve such problems rests on the idea that it will hit upon some set of optimal solutions which humans have not so far been able to. These solutions not only need to be effective, but also morally acceptable to all, and there is a limit on the time and resources that can be used to formulate and implement them. This is a tall order. If such solutions actually exist (and it is possible that they do not), they will presumably be fiendishly difficult to achieve even with the requisite intelligence. It is, of course, *possible* that our hypothetical, more achievable *human+ SI* will come up with a relatively cheap solution to climate change that is also relatively easy to implement in a timely manner. At the same time, there is a significant chance that this will not be the case.

At this juncture, two watered-down, but perhaps more plausible and worrying versions of the alignment problem re-emerge. On the one hand, as argued above, in order to create a SI within a reasonable timeframe, we would need to constrain its capacities. But how can we ensure that we do not, in doing so, inadvertently inhibit its intelligence in a way that precludes it hitting upon the requisite solutions? Moreover, given that our more probable human+ SI will not be maximally intelligent as envisioned by proponents, an aspect of the traditional alignment problem would apply to it.[9]

[9] This will again raise the problems relating to traditional alignment mentioned in Footnote 6.

Presumably, our human+ SI will understand human goals and interests, and this would limit the possibility that it will inadvertently cause human extinction or implement perverse solutions to our problems. However, it is not obvious that it will not pose a threat to humans. For one thing, humans often pose threats to one another. Society is replete with mechanisms to keep humans aligned with societal values (we make it in their interest to act pro-socially through punishments and rewards) and to deal with those who do not comply. Something similar may be needed for our SI. And even if we were able to create an aligned SI, we cannot assume that it will stay aligned. Dennett (2024), for example, convincingly argues that we cannot create entities with human-like intelligence without high degrees of autonomy. Simply put, greater intelligence requires greater autonomy. Autonomy enables entities to choose their own goals, to the extent that humans, for example, can override the constraining goals evolution has bestowed upon them. A SI capable of solving problems that humans cannot solve needs to have the ability to freely set goals, acquire relevant information, and assess possibilities and plan on the basis of that information in order to discover and assess possibilities and strategies that we have not been able to. As with human beings, such autonomy would necessarily give it the capacity to set goals that might, ultimately, be bad for human beings.

4. Conclusion

The case for SI, whether beneficial or not, is far from certain. It may be that SI is possible, achievable, and will necessarily be of immeasurable benefit to us. Nonetheless, we currently have no reason to be confident that SI is even possible, let alone imminent or inevitable. Moreover, even if SI were possible, we do not have good reason to think any possible SI will necessarily be benign or beneficial. Clearly, there are enough unquestioned assumptions and unanswered questions to point to the folly in pinning our hopes for resolving our biggest challenges on SI. Such folly would be exacerbated if we were to incur greater, known short-term risks in the hopes of bringing

about SI. Claims relating to the possibility and benefit of SI are extraordinary claims, which require extraordinary evidence in support of them. In the absence of such evidence, we have to assume that the odds that we will fail to create SI capable of resolving our challenges are much greater than the odds that we will succeed. And, given the magnitude of the threat of climate change, the cost of the very attempt in itself could be the extinction of humanity. Even if the cost were not the complete extinction of humanity, it is safe to assume that the quality of life for subsequent generations of humans will be severely impacted. All of this for a payoff that will almost certainly not be realised in time to be of much use to anyone. One may, on some consequentialist argument, want to claim that potential future benefits of SI will be of such magnitude that they will outweigh any current and near-future costs that humanity will occur, no matter how calamitous. However, even if one were to grant the consequentialist ethical framework employed here (which one need not do), the currently foreseen future benefits seem to rest on a series of assumptions that border on wishful thinking rather than on any scientifically-derived insights. We have no good reason to believe that any current or future generations will benefit from the risky gamble of accelerating the rate of AI development to the point of exacerbating climate change, which means that the magnitude of the payoff for future generations becomes a moot point. The risk is simply unacceptable.

References

Altman, S. (2024, 23 September 2024). The Intelligence Age. https://ia.samaltman.com/

Andrews, K. (2010). Animal cognition. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Annual Meeting of the Association for Computational Linguistics,

Bostrom, N. (2006). How Long before Superintelligence? *Linguistic and Philosophical Investigations*, *5*(1), 11-30.

This is a draft, pre-peer review copy and not the finalized manuscript. Please cite the final version, which will appear in the book, *Computers with Salaries and Cemeteries: AI Ethics from Industry to Philosophy to Science Fiction* (forthcoming)

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, *22*(2), 71-85. https://doi.org/10.1007/s11023-012-9281-3

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Cartmill, E. A. (2023). Overcoming bias in the comparison of human language and animal communication. *Proceedings of the National Academy of Sciences*, *120*(47), e2218799120. https://doi.org/doi:10.1073/pnas.2218799120

Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, *17*(9-10), 9 - 10.

Davis, E. (2015). Ethical guidelines for a superintelligence. *Artificial Intelligence*, *220*, 121-124. https://doi.org/https://doi.org/10.1016/j.artint.2014.12.003

Dennett, D. C. (1995). *Brainchildren: Essays on Designing Minds*. MIT Press.

Dennett, D. C. (2000). Kinds of Mind. *Mind*, *109*(436), 883-890.

Dennett, D. C. (2022). A Route to Intelligence: Oversimplify and Self-monitor. In S. Wuppuluri & I. Stewart (Eds.), *From Electrons to Elephants and Elections: Exploring the Role of Content and Context* (pp. 587-595). Springer International Publishing. https://doi.org/10.1007/978-3-030-92192-7_32

Dennett, D. C. (2024). We are all cherry-pickers. In A. Strasser (Ed.), *Anna's AI Anthology: How to live with intelligent machines?* (Vol. 9, pp. 15-30). xenomoi Verlag.

Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(5), 1-23.

Floridi, L. (2023). AI as agency without intelligence: on ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15.

Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In *Advances in computers* (Vol. 6, pp. 31-88). Elsevier.

Kant, I. (1908). Critique of pure reason. 1781. *Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin*, 370-456.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Viking.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

LeCun, Y., & Courant. (2022). A Path Towards Autonomous Machine Intelligence Version 0.9.2, 2022-06-27.

Legg, S., & Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, *17*(4), 391-444. https://doi.org/10.1007/s11023-007-9079-x

Lenat, D. B., & Marcus, G. (2023). Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc. *ArXiv*, *abs/2308.04445*.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517-540. https://doi.org/10.1016/j.tics.2024.01.011

Mandelbaum, E. (2022). Everything and More: The Prospects of Whole Brain Emulation. *Journal of Philosophy*, *119*(8), 444-459.

This is a draft, pre-peer review copy and not the finalized manuscript. Please cite the final version, which will appear in the book, *Computers with Salaries and Cemeteries: AI Ethics from Industry to Philosophy to Science Fiction* (forthcoming)

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books.

Mitchell, M. (2024). Debates on the nature of artificial general intelligence. *Science*, *383*(6689), eado7069. https://doi.org/doi:10.1126/science.ado7069

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 553-571). Springer.

Müller, V. C., & Cannon, M. (2021). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, *35*(1), 25-36.

Niemeyer, K. V., Lakshmi. (2024, 6 October 2024). Former Google CEO Eric Schmidt says we should go all in on building AI data centers because 'we are never going to meet our climate goals anyway'. *Business Insider*. https://www.businessinsider.com/eric-schmidt-google-ai-data-centers-energy-climate-goals-2024-10?op=1

Schwitzgebel, E. P., Jeremy. (2024). The Copernican Argument for Alien Consciousness; The Mimicry Argument Against Robot Consciousness. In. Unpublished.

Spelke, E. S. (2000). Core knowledge. *American psychologist*, *55*(11), 1233.

Spelke, E. S. (2022). Beyond Core Knowledge. In *What Babies Know: Core Knowledge and Composition Volume 1* (pp. 0). Oxford University Press. https://doi.org/10.1093/oso/9780190618247.003.0010

Thorstad, D. (2024). Against the singularity hypothesis. *Philosophical Studies*. https://doi.org/10.1007/s11098-024-02143-5

Tye, M. (2016). *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* Oxford University Press USA.

Vold, K., & Harris, D. R. (2023). How does Artificial Intelligence Pose an Existential Risk? In C. Véliz (Ed.), *The Oxford Handbook of Digital Ethics*. Oxford University Press.