Artificial Intelligence and Philosophical Progress

Graham Clay (grahamclay.com) & Caleb Ontiveros (calebontiveros.com)

March 25, 2021

The transformative power of artificial intelligence (AI) is coming to philosophy—the only question is whether philosophers will harness it. In this paper, we argue that the application of AI tools to philosophy could have a transformative impact on the field comparable to the advent of writing, and that it is likely that philosophical progress will significantly increase as a consequence of AI. The role of philosophers in this story is not merely to use AI but also to help develop it and theorize about it. In fact, we argue that philosophers have an obligation to spend significant effort in doing so.

"We live during the hinge of history. Given the scientific and technological discoveries of the last two centuries, the world has never changed as fast. We shall soon have even greater powers to transform, not only our surroundings, but ourselves and our successors."[1]

1. Introduction

The transformative power of artificial intelligence (AI) is coming to philosophy—the only question is whether philosophers will harness it. In this paper, we argue that the application of AI tools to philosophy could have a transformative impact on the field comparable to the advent of writing, and that it is likely that philosophical progress will significantly increase as a consequence of AI. The role of philosophers in this story is not merely to use AI but also to help

[1] Derek Parfit, *On What Matters* (Oxford: Oxford University Press, 2011), 616.

develop it and theorize about it. In fact, we argue that philosophers have an obligation to spend significant effort in doing so.

The structure of this paper is as follows. In section 2, we describe various forms of AI and the ways in which AI tools could be used by philosophers. After presenting one viable way of defining philosophical progress, we then illustrate how these tools would likely lead to significant progress of this sort. Our focus is twofold: we argue that philosophical progress would increase as a function of individuals using these tools and as a function of their institutional impact. In section 3, we argue that AI will be philosophically relevant soon—in that it can significantly assist philosophers within a reasonable timeframe—and that the degree of philosophical relevance depends on the involvement of philosophers. Finally, in section 4, we argue that philosophers are obligated to aim for philosophical progress. Given this obligation, philosophers are obligated to theorize about and develop philosophy-specific AI *now*, and utilize it as soon as it is available.

2. How Artificial Intelligence Would Help Philosophy Progress

There are many different ways to define philosophical progress, but one common way is in terms of truth. It is commonly argued that it is the truth of the views of philosophers that matters, such that if philosophers' true representations (factive epistemic states like knowledge) or approximately true representations (intermediate states that are neither ignorance nor knowledge) of philosophical propositions increase over a period, then philosophy progresses through that period.[2] This (partial) definition accounts for philosophers' representations of propositions ranging from *eating animal products is wrong* to *there is a distinction between metaphysical grounding and supervenience*.

[2] For a recent survey of the literature and a defense of a definition that incorporates our definition as a part, see Lewis D. Ross, "How Intellectual Communities Progress," *Episteme* (2020).

In this section, we will argue that AI tools would increase philosophical progress if progress were defined in this broadly epistemic way. However, we must emphasize that we are *not* committed to this epistemic definition being *the only plausible way* to define philosophical progress. Perhaps there are other equally viable ways to define philosophical progress. Yet, whether or not there are other varieties of progress, epistemic progress of this sort clearly matters in the morally significant way that we need for our argument to go through—that is, for us to establish, via the principles discussed in the later sections of this paper, that philosophers should theorize about and develop philosophy-specific AI now, and that they should utilize it as soon as it is available. So, in what follows, we will operate with this definition presumed in order to establish our position and reveal a defensible path to our conclusion, but we maintain that AI would increase philosophical progress in many other morally significant ways, too, and that these other effects constitute other plausible paths to our conclusion.

2.1 - AI and AI tools

Writing has significantly increased philosophical progress. Indeed, it is hard to overstate the impact of writing on philosophy. We could not be footnotes to Plato were it not for this powerful tool. Imagine waking up and learning that, due to a freak cosmic accident, all books, journal articles, notebooks, blogs, and the like had vanished or been destroyed. In such a scenario, philosophy would be made seriously worse off. Present philosophers would instantly suffer a severe loss and future philosophers would be impoverished as a consequence.

The advent of writing freed philosophers from being solely dependent on their own memories and oral methods of recollection. It enabled philosophers to interact with other thinkers across time, diminishing the contingent influences of time and space, thereby improving

the transmission of ideas. Philosophers were able to learn about others' approaches to philosophy, which in turn aided them in their own methodologies.

It is our position that AI would provide a suite of tools that can play a similar role for philosophy. But what exactly is AI? In this subsection, we answer this question and provide some insight into the sort of AI tools that would transform philosophy.

There are, very roughly, two kinds of AI: machine reasoning and machine learning systems. Machine reasoning systems are composed of knowledge bases of sentences, inference rules, and operations on them.[3] For example, you could have a program with the following sentences:

P: It is possible for there to be a physical duplicate of me that is not conscious.

Q: If P, then eliminative materialism is false.

This system could include among its inference rules the inference rules of classical first order logic and the ability to form new sentences by applying those inference rules. If you designate P and Q as true in this system, then the following sentence would be outputted:

R: So, eliminative materialism is false.

Although it is simple, such a system would be a machine reasoning system.

Another kind of AI is a machine learning system. Such a system works by ingesting a large amount of data and learning to make accurate predictions from patterns contained in it. In a modern context, machine learning is implemented using deep learning and related techniques.[4] Two especially impressive machine learning systems are AlphaGo and GPT-3. AlphaGo was the

[3] The contemporary system taking this approach is Cyc (for discussion, see Douglas B. Lenat, "Cyc: A Large-Scale Investment in Knowledge Infrastructure," *Communications of the ACM* 38, 11, (1995).

[4] Technical background on these techniques is not required for understanding this essay, but for those interested, an introduction is found in Cameron Buckner, "Deep learning: A philosophical introduction," *Philosophy Compass* 14, 10, (2019).

first AI system to beat a professional human player in a full-sized game of Go, beating the European champion 5 games to 0.[5] GPT-3 is a natural language model that is able, with minimal input, to produce paragraphs of content responding to queries about fictional characters, music, politics, and even philosophy.[6] Its responses are novel but often indistinguishable from those that would be given by a human responding to the same queries.

One can think of the first kind of system—machine reasoning AI—as a deductive and symbolic reasoner and the second—machine learning AI—as learning and implementing statistical rules about the relationships between entities like words.[7] Presently, our concern is with how such systems could increase philosophical progress. There are many ways in which they could do so. These systems have vastly superior computational speed. GPT-3 can ingest and produce paragraphs of text at a much faster rate than philosophers. (It has been trained on *hundreds of billions* of words.) Likewise, these systems have much greater storage space than humans. A given human has access to only a portion of the Stanford Encyclopedia of Philosophy at a given time, while an AI system could have access to all of it, in addition to all of the articles in the bibliographies it contains.

Below are some of the most promising philosophical applications of AI, ordered from the mundane to the more speculative:

- *Recommendation*: locating and suggesting useful content.

---

[5] Silver, D., Huang, A., Maddison, C., et al. "Mastering the game of Go with deep neural networks and tree search," *Nature* 529, (2016): 484–489.
[6] Tom B. Brown, et. al. "Language Models Are Few-Shot Learners," <https://arxiv.org/abs/2005.14165>, (2020).
[7] These do not exhaust possible AI implementations. Research into Neurosymbolic AI aims to create hybrids of the two systems. See, e.g., Artur D'Avila Garcez et al., "Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning," <https://arxiv.org/abs/1905.06088>, (2019), and Artur D'Avila Garcez & Luis C. Lamb, "Neurosymbolic AI: The 3rd Wave," < https://www.arxiv-vanity.com/papers/2012.05876/>, (2020).

- *Synthesis*: summarizing existing work and ensuring that current encyclopedias are up to date.
- *Systematizing*: relating philosophical propositions and positions.
- *Simulation*: providing germane contributions from the standpoint of a simulated philosopher or a simulated believer of a given position.
- *Formalizing*: transforming common language statements into formal logic.
- *Reasoning*: reasoning through philosophical propositions in a way that is philosophically useful.

Recommendation tools would locate and suggest useful papers, books, lectures, and other content relevant to the interests and activities of users. While there are existing tools that do this without the use of AI, such as the Stanford Encyclopedia of Philosophy and PhilPapers, Recommendation tools driven by AI would be superior in some respects. Consider how students often become overwhelmed and are unable to locate philosophical content suitable for their level of expertise when confronted with the dizzying array of information contained in these encyclopedias and databases. Currently, domain experts like professors are tasked with answering student questions about the relevant literature in ways suited to the expertise of the student. Recommendation tools could either handle this task entirely or help professors do so. On both options, Recommendation tools would make philosophers faster at answering student queries or simply enable them to dedicate their time to research instead. Likewise, professors could use Recommendation tools that have access to their paper drafts—much like spell-checking tools do already—in order to find citations and pertinent literature in real time.

Synthesis tools would summarize existing philosophical literature. Such tools would help keep digital encyclopedias' summaries of extant literature up to date with the most recent

advances. As such tools' capabilities to summarize a philosophical domain increase, we would expect that philosophers' ability to understand and productively reason about that philosophical domain to increase as well.

Systematizing tools would relate philosophical propositions to one another. Currently, there are several websites, including PhilPapers, that relate philosophical papers by citations and content. An AI-driven Systematizing tool would do the same thing but for propositions and collections of them (i.e., philosophical positions or systems). Many different domains in philosophy are interconnected in interesting ways. Positions in philosophy of language, for instance, support particular views in metaethics and metaphysics.[8] While philosophers face storage and computational constraints when determining whether and how two given positions or propositions are connected, an AI system would not. Propositions might be related by similarity at the content level or by more interesting structural properties like support or conditional likelihood (e.g., P is supported by Q, or R is more likely given S). Such a tool could improve philosophers' ability to recognize relevant connections between propositions and, in turn, the arguments they compose.

An additional way that such a tool may assist philosophers is by helping them generate novel arguments and counterexamples. For many of our novel thoughts, we arrive at them by relating two concepts in a new way. For example, we might relate the concepts of *justice* and *effective altruism* by combining them into the idea of *effective justice*.[9] Through the use of a Systematizing tool, philosophers would be able to filter through potential relations between

[8] For examples, consider Saul Kripke, *Naming and Necessity* (Oxford: Basil Blackwell, 1980); John MacFarlane, *Relative Truth and its Applications* (Oxford: Oxford University Press, 2016); and Timothy Williamson, *The Philosophy of Philosophy* (Oxford: Wiley Blackwell, 2008). For discussion of this interconnectivity in relation to progress, see Frank Jackson, "Philosophy, Progress, and Identity," in R. Blackford & D. Broderick (eds.): *Philosophy's Future: The Problem of Philosophical Progress* (Oxford: Wiley Blackwell, 2017),227-239.
[9] Roger Crisp & Theron Pummer, "Effective Justice," *Journal of Moral Philosophy* 17 (2020).

philosophical positions in order to select and refine relations they may not have thought of otherwise. Although Recommendation and Synthesis tools would be powerful, Systematizing tools could be truly transformative to philosophy. We will consider them in more detail below, when discussing how philosophers can contribute to the development of AI tools.

Simulation tools that simulate other philosophical positions or philosophers would aid philosophers' problem-solving ability in the same way that having a capable interlocutor does. But to be clear, we are not here concerned with emulating or simulating a philosopher in the sense that the AI system would qualify as a philosopher (or as a mind or as having a mind). Rather, we are concerned with a tool that would answer questions like "What would Philippa Foot say about this?" or "What would a Platonist say about this?" with reasonable answers. Such a tool would be useful even if there was no sense in which it understands anything.

As a contemporary example of this, consider GPT-3's response to the query: "How does a theist respond to the problem of evil?":

> I am a theist, and I believe in the existence of an omniscient, omnipotent, just deity named God. You would think this belief is incompatible with evil; but it isn't.
>
> In my view, the existence of evil is not a refutation of God's existence; rather it is one form of evidence for His existence.
>
> The existence of evil is what gives us the power to distinguish good and evil. If there were no such thing as evil, then we would have no way to tell right from wrong, or moral from immoral.
>
> It is a well-known fact that you cannot make moral judgments without having some concept of evil. Thus, if there were no such thing as evil, it would be impossible to make any kind of moral judgment.

> The existence of evil also allows us to detect and appreciate good. If there were no such thing as pain, we would have no way to know what it is like to feel good; we might think that all pleasure was normal.[10]

Although current technology is not at the point where it can reliably answer philosophical queries, there have been compelling advances over the past few years.[11] Likely, such tools will first be useful for teaching and philosophical training. However, over time, such simulations will become immensely useful for philosophical experts. Imagine being able to talk to any challenging interlocutor from the philosophical canon!

Formalizing tools would help test existing projects, such as the creation of deontic logics, by formalizing arguments and by showing philosophers different ways of doing so. They would improve the speed of philosophers in reasoning through arguments and could play an important role in philosophical training, much in the same way pedagogical tools dedicated to argument mapping do already. Such tools would also make other AI tools more useful by, say, formalizing the results of Simulation tools.

Subsequent to formalizing other tools' outputs, Formalizing tools could create inputs for *Reasoning* tools that reason through philosophical propositions and domains in a useful way. Think of Reasoning tools as similar to the simple machine reasoning program described above that included the following propositions:

P: It is possible for there to be a physical duplicate of me that is not conscious.

Q: If P, then eliminative materialism is false.

Philosophers (or another AI tool) could input hundreds or thousands of propositions

[10] This example was generated by the authors using GPT-3. It was the second response to the query. The first response to the query was coherent but denied that God was morally perfect (an unconventional theistic response).
[11] More research is required to determine how reliable GPT-3 is at answering philosophical queries coherently. We suspect that it varies considerably on the implementation and query text.

concerning eliminative materialism into the Reasoning tool. The tool would then be able to output consistent sets of propositions or—given certain further epistemological inference rules—propositions about what you should believe about these propositions given your beliefs in others. The ideal version of this tool would include billions of such propositions with the ability to apply hundreds of different logics to them so as to enable philosophers to map their consequences and other relations. The feasibility of this ideal is unclear, but exceptional gains could be realized with more limited versions, too. Such a tool would assist philosophers in reasoning through a much wider range of propositions and at a faster speed than they would have otherwise. They would be able to quickly craft a higher resolution picture of the relations between particular propositions and the status of different arguments involving them. A philosopher could, for instance, realize that an argument concerning eliminative materialism equivocates (by seeing the ways in which the Reasoning tool evaluates two premises sharing a term); uncover one of their conceptual confusions; or discover new relations and entailments between common arguments that had not been noticed or discussed before.

2.2 - The impact of AI on progress

On the epistemic definition of philosophical progress, if philosophers' true or approximately true representations of philosophical propositions increase over a period, then philosophy progresses through that period. There is disagreement about how exactly to cash out this sort of definition. Some argue that the relevant propositions are generally going to be new in the sense that philosophers have not considered them previously, whether in their era or at all.[12] Others focus on philosophers' representations of perennial propositions like *God exists*.[13] And while some

[12] For relevant discussion and a defense of a version of this claim, see Bryan Frances, "Extensive Philosophical Agreement and Progress," *Metaphilosophy* 48, 1-2 (2017): 47-57.
[13] See, e.g., David J. Chalmers, "Why isn't there more progress in philosophy?," *Philosophy* 90, 1 (2015): 3-31.

focus on individual philosophers' representations, others argue that it is the epistemic position of the "intellectual community" as a whole that is relevant.[14] On all such views, it is not sufficient for a single isolated philosopher to arrive at true representations of philosophical propositions for philosophy to progress, given that the relevant sort of progress is that of the field *as a whole*. This is the perspective we will adopt, although recall that we are open to there being other viable ways of defining progress.

We maintain that it is likely that AI would help philosophy progress in a significant way. Here is our argument for this claim:

P1. Philosophers who can use AI tools would be significantly better able to locate plausible alternatives to their positions, find literature relevant to their research, detect errors in their reasoning, generate novel philosophical thoughts, and justify philosophical propositions with arguments than those who cannot.

P2. Philosophers who can use AI tools would be significantly better able to fruitfully interact on a regular basis with their peers and construct institutions for transmitting and doing philosophical work.

P3. If P1 and P2, philosophers who can use AI tools would be significantly more likely to have more true or approximately true representations of philosophical propositions than philosophers who cannot.

C1. So, philosophers who can use AI tools would be significantly more likely to have more true or approximately true representations of philosophical propositions than philosophers who cannot.

P4. If philosophers' true or approximately true representations of philosophical

[14] For a discussion of the literature on this point, see Ross, "How Intellectual Communities Progress," 2020.

propositions increase over a period, then philosophy progresses through that period.

C2. So, it is likely that AI tools would help philosophy progress in a significant way.

The foundations for our argument for P1 are found in the preceding subsection. As we have discussed, philosophers who can use the aforementioned AI tools—from Recommendation to Reasoning tools—would be assisted in every aspect of their work, from teaching to research. This assistance would make them more productive and effective regardless of the task they seek to complete because *either* there would be AI tools that directly improve their ability to do the task in question *or* there would be AI tools that make them more efficient with their other tasks and thereby free up their time and energy for the task in question. Locating plausible alternatives to their positions, finding literature relevant to their research, detecting errors in their reasoning, generating novel philosophical thoughts, and justifying philosophical propositions with arguments are among the main tasks of philosophers, and they are tasks that AI tools would be especially useful in assisting with.

For instance, Systematizing tools would locate plausible alternatives to a philosophers' positions by showing relations between these positions and other propositions, thereby enabling them to efficiently locate implausible consequences of their views as well as nearby views that avoid these consequences. Simulation tools could be used to argue persuasively for these alternatives to show philosophers the sort of resistance they might meet in virtue of their positions. Likewise, Recommendation tools would find literature that would reveal considerations that a researching philosopher would not have thought of otherwise, and Synthesis tools would summarize this literature into a form that would save her time. Finally, Formalizing and Reasoning tools would convert philosophers' arguments into natural language premise-conclusion form and/or logical notation, which would in turn expose invalidities and

equivocations. By assisting with all of these tasks, AI tools would make philosophers significantly better at them, as well as the tasks they themselves compose, like justifying philosophical propositions with arguments. In effect, AI tools would provide an enhanced version of the environment in which philosophers already seek to immerse themselves with research assistants, academic conferences, and the like.

Of course, there are many other ways in which AI tools would positively affect philosophers beyond these examples. These examples are simply straightforward illustrations of the great promise of AI that we have chosen in part because they bear tight connections with the epistemic definition of philosophical progress. But, before we turn to this connection (as expressed by P3), we pause to note some of the *institutional* consequences of AI. While the preceding illustrations reveal impacts of AI on *individual* philosophers, we will now argue, per P2, that there would be significant institutional impacts of AI, too.

We maintain that it is likely that AI would help decrease philosophical disagreements over time *for the right reasons* by making philosophers significantly better able to fruitfully interact on a regular basis with their peers. AI would enable philosophers to communicate better with one another about alternatives to their positions, relevant literature, potential errors in their reasoning, and the arguments that they think best justify philosophical propositions, thereby enabling them to convince one another and generate consensus more effectively. AI would bring disparate areas of philosophy into contact, as well as diminish the biasing effects of prestige-driven journals by decreasing their relevance.[15]

[15] For relevant discussion of institutional barriers to progress that *we* are here arguing would be significantly improved by AI, see Jessica Wilson, "Three Barriers to Philosophical Progress," in R. Blackford & D. Broderick (eds.): *Philosophy's Future: The Problem of Philosophical Progress* (Oxford: Wiley Blackwell, 2017), 91-104.

For instance, possible AI systems would be able to reveal to each of their users the versions of the cosmological argument that the philosophical community as a whole believes best justify the proposition that God exists; the arguments philosophers think best justify the premises of these cosmological arguments; the AI-generated connections between these justifications; alternative justifications discovered by the AI; and so on. And since the user interfaces of these systems would be designed to facilitate direct communication between users as well (if desired), philosophers could engage with one another in a topical and efficient way without attending conferences or publishing in journals. In these ways and more, AI would lower the transaction costs typical of the philosophy profession, just as PhilPapers and PhilPeople have done so already, except to a greater degree. (Again, this is *not* to say that there are no other important aspects of philosophical practice that AI would impact beyond these illustrative examples.)

P3 links these institutional and the aforementioned individual impacts of AI to a significantly increased likelihood of having more true or approximately true representations of philosophical propositions. With regards to the *individual* impacts of AI, we maintain that philosophers' ability to locate plausible alternatives to their positions, find literature relevant to their research, detect errors in their reasoning, and justify their positions with arguments could not fail to positively correlate with their likelihood to have true or approximately true representations of philosophical propositions. The senses of 'plausible,' 'relevant', 'error', and 'justify' at issue here are primarily epistemic. The extent to which an alternative is plausible is, in large part, the extent to which it is a contender for being true; the extent to which literature is relevant to research is, in large part, the extent to which the literature helps the researcher discover or argue for the truth; the extent to which a property of a chain of reasoning is an error

is, in large part, the extent to which it leads to falsehood; and the extent to which philosophical arguments justify their conclusions *just is* the extent to which their truth makes their conclusions more likely to be true or approximately true, at least in nearly every case. All else equal, philosophers who are better able to do these tasks are significantly more likely to have true or approximately true representations, since their representations of philosophical propositions will be at least somewhat influenced by their ability to do these tasks. (This is made all the more plausible by the fact that philosophers who are better able to do these tasks will be better able to judge the extent to which they have true or approximately true representations of philosophical propositions.)

With regards to the *institutional* impacts of AI, we maintain that philosophers who are better able to fruitfully interact on a regular basis with their peers are significantly more likely to have true or approximately true representations of philosophical propositions than philosophers who are not. AI tools can improve philosophical practice *synchronically* by improving communications within the contemporary philosophical community and *diachronically* by improving the philosophical communities' ability to communicate across generations. The connections between increasing philosophers' fruitful interactions, increasing consensus (for the right reasons), and increasing the likelihood of philosophers having true or approximately true representations are tight. Since it is not possible for both a philosophical proposition and its negation to be true, if the amount of true or approximately true representations of philosophical propositions increases over time across all philosophers, then typically there will be a "convergence to the truth" amongst philosophers.[16] More importantly, though, the relationship

[16] For discussion, see Chalmers, "Why isn't there more progress in philosophy?". In order for the number of true or approximately true representations to increase across philosophers, representations need to be distributed in such a way that these representations are not concentrated in a small subset of philosophers. Given the historical record of convergence, and the interrelatedness of many philosophical issues, we believe that this sort of distribution can

goes the other way, too, *from* convergence *to* truth, at least in the case of convergence *for the right reasons*. If there is widespread and persistent disagreement through a period, and if all of the philosophers are roughly equal in ability, not subject to significant cognitive or affective biases, able to communicate with one another, and so on, then it is likely that the quantity of philosophers' true or approximately true representations of philosophical propositions has not significantly increased during it. If, by contrast, such a community of philosophers converge (in that they represent the same propositions as true), then they converge for the right reasons and are consequently more likely to be in possession of the truth.[17] Since convergence of this sort is made significantly more likely by significantly increasing the ability of philosophers to fruitfully interact on a regular basis with their peers—which is an impact of AI—it follows that this would make philosophers significantly more likely to have more true or approximately true representations of philosophical propositions than those who are not.

Note that P3's defensibility does not turn on the sorts of philosophical propositions at issue. There is no significant difference between new and perennial philosophical propositions that affects the extent to which philosophers are able to justify them via arguments. As discussed previously, perennial philosophical propositions are those like *God exists*; *we know about the external world*; and *we have free will*. New philosophical propositions are defined as new relative to an era or a time period, but recent new propositions include *the concept of supervenience is required to distinguish physicalism from the alternatives*; *knowledge entails*

typically be presumed. Clearly, though, there are periods where true or approximately true representations could increase in virtue of new discoveries but convergence would decrease in virtue of the inertia of old falsehoods.

[17] Perhaps it is for this reason that many of those who argue that there is widespread and persistent disagreement in philosophy, as well as those who object, are concerned with the topic because of its consequences for progress. See, e.g., Chalmers, "Why isn't there more progress in philosophy?"; Daniel Stoljar, *Philosophical Progress: In Defense of Reasonable Optimism* (Oxford: Oxford University Press, 2017); and Ward E. Jones, "Philosophy, Progress, and Identity," in R. Blackford & D. Broderick (eds.): *Philosophy's Future: The Problem of Philosophical Progress* (Oxford: Wiley Blackwell, 2017), 227-239.

safety; and *how justified one is in one's political convictions depends on how close one is to achieving reflective equilibrium*. There is no deep distinction between these two kinds of propositions.

With P1, P2, P3, and the epistemic definition of philosophical progress (P4) in hand, it follows that it is likely that AI would help philosophy progress in a significant way (C2), whether philosophical progress is defined in terms of the truth or the approximate truth of philosophers' representations of philosophical propositions, whether the relevant propositions are new or perennial, and whether philosophers are treated as a collection of individuals or as an intellectual community.

3. How Philosophers Can Make AI Philosophically Useful

3.1 - The near-term feasibility of philosophically relevant AI

We have argued that it is likely that AI *would* help philosophy progress in a significant way, both through individual and institutional impacts. Whether it *will* is another matter. The feasibility of philosophically relevant AI depends on factors like the technological capabilities and developmental trajectory of AI, philosophers' willingness to contribute to the development of AI, and their willingness to utilize it. In this subsection, we will discuss the first of these factors and argue for the following premise:

> P5. Within the next decade, it is likely that the technological capabilities of AI will be sufficiently sophisticated to be philosophically relevant.

Philosophically relevant AI is AI that can significantly assist philosophers, including in the ways described above. There is uncertainty about when philosophically relevant AI will arrive. However, we believe that the likelihood is high enough to justify doing philosophical work *now* in order to best bring about and prepare for its arrival so as to maximize philosophical

progress. We will give three arguments for this claim about likelihood: the *argument from compute*, the *argument from current progress*, and the *argument from expert judgment*. Because this paper is not focused on forecasting, we will not belabor the details of the arguments or give quantitative estimates, such as confidence intervals. Nonetheless, completely ignoring the question of feasibility risks resting the overall thesis of this paper on what might seem to some as a fantasy.

The *argument from current progress* is simple. AI has improved rapidly in recent years. Since it is likely that this trend will continue, it is likely that AI will be capable of significantly assisting philosophers in a reasonable timeframe. By "reasonable timeframe," we mean that there is a good chance that some important AI tools—namely, Recommendation, Synthesis, and Systematizing tools—will be realized within the next decade. (We will set the more speculative tools like Formalizing and Reasoning tools aside for now, but we are confident that they have a good chance of being realized by 2100.) A good chance is one that is high enough to justify philosophers acting as we recommend in the next subsection. Given that such tools would help philosophy progress in a significant way, the likelihood that they come to fruition does not need to be exceptionally high for considerations about it to be action-guiding in this sense.

All things considered, current technological trends are weak but sufficient evidence for continued progress. Here are several notable recent successes:

- AlexNet wins the ImageNet Large-Scale Visual Recognition Challenge with a 15% error rate in 2012.
- AlphaGo defeats Lee Sedol, one of the best go players in the world in 2016.
- AlphaGoZero defeats AlphaGo without training on any human games in 2017.

- AlphaZero, a generalized version of AlphaGoZero, is able to win games of chess against one of the best AI programs, Stockfish, *and* beat AlphaGoZero in 2018.
- OpenAI 5 defeats the world-champion team OG in Dota 2, a multiplayer video game, in 2019.[18]
- OpenAI's GPT-2 and GPT-3 are released in 2019 and 2020.

Illustrating the promise of GPT-3, David Chalmers used it to produce exchanges like the following (note that the following is output from GPT-3 itself, *not* Chalmers):

> [GPT-3] might be conscious, but we don't know yet.
>
> I'll give you an example. Suppose I told you that there was a man named Bob who lived in your town. You could tell me a lot about him by looking at his house. If he had two cars, you'd have more information than if he didn't. He might own a boat, so you'd have even more information about him. You might know that he was married with two children, and that would be a lot of information about him as well.
>
> All of this information is structural, and as I said, it all tells me very little about Bob's state of mind. All I know is that he owns a house, has two cars, and is married. But if I already know these things, and I see that his house has been broken into, I can safely say that he is a mind with fears and concerns. I can say that he is a mind with a certain past, certain present, and certain future concerns.
>
> The exact same thing can be said about GPT-3. It has some computational processing, and it instantiates certain functional properties. But as for what it's like to be GPT-3, we have no idea. It could be anything.[19]

---

[18] Christopher Berner et al. "Dota 2 with Large Scale Deep Reinforcement Learning," OpenAI (2019), <https://cdn.openai.com/dota-2.pdf>.

[19] Justin Weinberg, "Philosophers On GPT-3 (Updated with Replies by GPT-3)," *Daily Nous*, <dailynous.com/2020/07/30/philosophers-gpt-3/>, 30 July 2020.

This is a coherent response that gestures at considerations that are commonly discussed in the literature. Indeed, Chalmers notes that he agrees with the key point that "structure can be evidence for consciousness even if it does not constitute consciousness." With additional development, it is reasonable to expect that GPT-*n* will improve at coherently and usefully answering philosophical queries. And there are other AI-driven classification and summarization tools that get close to what is required to count as Synthesis and Systematizing tools.[20] If progress continues at the current rate—and it is likely that it will, given the momentum and success of the industry—there is a good chance that such tools will be philosophically useful within the next decade.

The *argument from compute* is that since it is likely we will see a continued increase in computing power in the next decade, it is likely that we will see advances in AI over the same time period sufficient to be philosophically relevant.[21] Compute power refers to the number of operations per second achievable by computer hardware systems. It has been increasing rapidly over the past few decades, a trend which is a consequence of Moore's law.[22] The success of deep learning depends on the amount of computing power. Significant commercial and scientific success has been enabled by recent increases in computing power. For instance, billion parameter models, like GPT-3, would have been infeasible to train in the same timeframe a decade ago.[23] There have been continual and significant algorithmic advances over the last

---

[20] Jingqing Zhang et al., "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," <https://arxiv.org/abs/1912.08777>, (2019).

[21] For an introduction, see Dario Amodei & Danny Hernandez, "AI and Compute," OpenAI, <openai.com/blog/ai-and-compute>, 2018.

[22] Gordon Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, 38, 8, (1965): 114-117. Moore's law states that the number of transistors in a chip doubles every two years. This enables operations to be performed faster and faster. There is some dispute over whether this law has continued to hold since the 2000s, but it is uncontroversial that computing power is doubling every few years. See José Luis Ricón, "Progress in semiconductors, or Moore's law is not dead yet," *Nintil*, <https://nintil.com/progress-semicon/>, 2020.

[23] GPT-3 was trained on V100 GPUs which were introduced in 2017. See Brown et al., "Language Models are Few-Shot Learners."

decade that should not be underestimated, but deep learning models need data to train on, and having more computing power has allowed the models to train quicker and on more data (and cost less as a consequence).[24] We expect this trend to continue, powering more advanced AI tools like the ones discussed above.[25]

The *argument from expert judgment* is that because relevant experts take seriously the likelihood of AI undergoing significant development over the next decade, philosophers should believe that it is likely to be philosophically relevant in this time period. In the largest study to date, when asked the chance that AI will be "able to accomplish every task better and more cheaply than human workers," the average AI expert estimated a 50% chance by 2061 and a 10% chance by 2025.[26] (You read that right—*every* task!) Moreover, many researchers are increasingly worried about the dangers posed by artificial general intelligences (AGIs).[27] The kind of AI tools we are concerned with here are *not* AGIs, and it is a separate question how these tools and AGIs are related. Yet, the fact that there is significant concern about the creation of human level (and greater) AGI lends weight to the view that the tools specified above are feasible since they are vastly less complex and technologically demanding.

These three arguments justify thinking that within the next decade, it is likely that the technological capabilities of AI will be sufficiently sophisticated to be philosophically relevant (P5). It is worth noting that we are *not* arguing that AI will need to obtain anything like artificial

---

[24] Hernandez & Brown, "Measuring the Algorithmic Efficiency of Neural Networks."
[25] Suggestively, Hans Morevec has argued that computers will reach the computational power of the human brain in the 2020s. See Hans Morevac, "When Will Computer Hardware Match The Human Brain," *Journal of Evolution and Technology* 1, (1998).
[26] Katja Grace et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," *Journal of Artificial Intelligence Research*, 62, (2018): 729-754.
[27] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2017) and Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books, 2020).

general intelligence or superhuman intelligence, as the tools that we suggested above can be made without any such capabilities.

3.2 - Philosophers' role

One way for it to be *more* likely that progress in AI development is philosophically useful is if philosophers play a role in theorizing about and developing it. In this subsection, we will motivate the following premise:

P6. Unless philosophers theorize about and develop philosophy-specific AI, it is likely that philosophically relevant AI will not be as useful.

Let us make this concrete with a specific philosophical tool: *Systematizing*. This tool would encode relations between philosophical propositions, such as support or conditional likelihood (e.g., P is supported by Q, or R is more likely given S). Supposing that such a tool would generate its own encoding of these relations or acquire them from another tool, there are two ways that philosophers' input would significantly further its development: assessment and theorizing.

With regards to the first, philosophers would be needed in order to assess how well such a system is performing. Are the propositions truly related in the relevant way? If the system says that proposition T supports proposition U, is that correct? In a toy model of AI, a human hands off problems to an AI which it would solve without human involvement. This model is unrealistic given AI's current capabilities, especially in complex domains like philosophy. A more realistic way to understand these tools is as partially constitutive of human-AI hybrids, in the sense that a human must be in the loop in order for the tool to be successful. Examples of such hybrids in other domains include:

- AI that transcribes text to audio which a human must edit in order to assess the accuracy of the transcription.
- AI that transcribes audio when it has high confidence of accuracy but passes on segments of the audio about which it has a low confidence to a human transcriber.
- AI (or several) that plays chess together with a human by suggesting moves to the human who then chooses what to play from these options.

Like these non-philosophical AI tools, we should expect AI-driven Systematizing tools to require human oversight and feedback. Such feedback work might involve working with computer scientists to score the relations that the Systematizing tool produces and learning how to use the system in a way that produces the most philosophically valuable relations.

Extending this to the Reasoning tool, we can imagine a human-AI hybrid version of this tool that takes submissions in a manner that journals do, but instead of submitting a paper, philosophers would submit a set of propositions and inference rules. Other philosophers would review these submissions and either request edits or reject them if they are too far from being sound. Accepted sets would be added to the system and integrated with other accepted sets. Editors and reviewers would evaluate any conflicts in the system, call for special issues in cases of philosophically fruitful conflicts or consequences, and so on. Over time, this system would become an accurate representation of philosophical knowledge in a way that parallels current journals. Yet it would improve on current journals in that the Reasoning tool would be able to reason over this knowledge, integrate it, and expand on it by attempting to justify additional propositions (or noting plausible justifications given its dataset). Philosophers would be able to improve the output of the system, which would in turn improve philosopher's skill, thereby achieving modest positive feedback loops. (Note that in addition to being a model of how

philosophers can play a crucial role in the development of philosophical AI tools, this specific tool is an example of a tool which satisfies P3 by improving philosophers' institutional capabilities and thereby contributing to progress.)

The second way that philosophers would significantly further the development of philosophically relevant AI is by theorizing about the epistemic value of such systems. It is one thing to be confident that a specific relation between propositions is true, but it is another to be confident that the system can infer true relations over time in many different domains and circumstances. The challenge here is to analyze and evaluate what it takes for such AI tools to be epistemically reliable and useful, how such AI tools could be improved, and so on. This problem becomes more salient when it is unclear whether, given the internals of its program, the AI understands or knows anything in any relevant sense (which appears to be the case for the best machine learning systems).

There is already a concern among computer scientists that deep learning and similar algorithms act as "black boxes" in that there are no accessible or interpretable reasons for their outputs.[28] This is a problem that philosophers are well posed to assist with. It is important to be justified in believing that the outputs of such systems are epistemically justified, regardless of which of the previously enumerated types they are. Having such justification requires being able to assess the process that created the relevant output. However, what this amounts to and how it can be done is not straightforward in the current AI models, whether they are identifying images, translating text, or predicting criminal activity. There is exciting conceptual work here that will

[28] For extensive treatment of this issue, see Kathleen A. Creel, "Transparency in Complex Computational Systems," *Philosophy of Science* 87, 4, (2020). For different uses of the term *interpretability*, see Zachary Lipton, "The Mythos of Model Interpretability," <https://arxiv.org/pdf/1606.03490.pdf>, 2016.

be useful for creating philosophically relevant AI (and better AI systems in general).[29] This work will require familiarity with epistemology, but also the ability to reason through the internals of the tools.

We do not doubt that AI tools will be significantly epistemically valuable within a reasonable timeframe. We have suggested applications, but more work would ideally be done by philosophers to reflect on them and others. This philosophical and practical work will render AI tools more philosophically useful than they would otherwise be.

4. The Obligation to Maximize Progress

The arguments of the preceding sections established the following claims:

C2. So, it is likely that AI would help philosophy progress in a significant way.

P5. Within the next decade, it is likely that the technological capabilities of AI will be sufficiently sophisticated to be philosophically relevant.

P6. Unless philosophers theorize about and develop philosophy-specific AI, it is likely that philosophically relevant AI will not be as useful.

In section 2, we established C2, while in section 3 we established P5 and P6. In this section, we will argue for the following moral claim:

C3. Philosophers should theorize about and develop philosophy-specific AI *now*, and they should utilize it as soon as it is available.

C3 follows from the conjunction of the preceding claims and this new moral claim:

P7. Philosophers should seek to maximize philosophical progress.

[29] See Creel, "Transparency in Complex Computational Systems" for a conceptual discussion of interpretability and, for a discussion of this issue in the context of neuroscience, see Mazviita Chirimuuta, "Prediction versus understanding in computationally enhanced neuroscience," *Synthese*, (2020).

As before, here we are relying upon a broadly epistemic notion of philosophical progress. Philosophy progresses if philosophers have more true or approximately true representations of philosophical propositions. So, P7 should be understood as asserting that philosophers should attempt to maximize the amount of true or approximately true representations of philosophical propositions had by philosophers. This is a claim about what one of the universal ends or purposes of philosophizing is—namely, that it is to try to find and promote philosophical truths amongst philosophers. Of course, philosophy has many ends or purposes, but we claim that this is one of them, and it is universal in that, *prima facie*, it applies to all philosophers, or so we will argue.

If philosophers seek to maximize philosophical progress, they are attempting to bring about something that has value. If, for example, knowledge was intrinsically valuable and if knowing entails having true representations of what one knows, then philosophical progress as we have defined it would be intrinsically valuable. Philosophical progress would also be intrinsically valuable if understanding, wisdom, or other similar goods were intrinsically valuable. On this picture, to the extent that philosophers attempt to maximize philosophical progress, they attempt to increase this variety of intrinsic value as much as they can. Whether they are increasing their own true or approximately true representations of philosophical propositions or they are increasing those of other philosophers, they are increasing the quantity of intrinsically valuable things.

While we are not committed to the view that philosophical progress has intrinsic value, we are committed to it having extrinsic value. When philosophy progresses, the outputs of philosophers become, on average, more truthful. And when the representations of philosophers get closer to the truth, so do the propositions they affirm and express in speech and in writing.

Academic research outputs of philosophers like journal articles and books contain more truths. Popular research outputs like op-eds, magazine articles, and essays in intellectual magazines contain more truths. These truths have ripple effects across the broader intellectual economy. Given how much the academic research in year *n* depends on the academic research of year *n-1*, the momentum of academic research increases in the direction of truth. The multiplicative effects over decades are massive. And the non-philosophers who read the popular research outputs are similarly affected. The claims and arguments they discuss and defend on the basis of what they read contain more truths. Policymakers influenced by these outputs, by constituents influenced by them, or by direct interaction with philosophers are in possession of the truth more often than had philosophy not progressed.[30] Whether the truths in question concerned the epistemological aspects of climate science skepticism, the role of religion in the public sphere, the nature of economic justice, or, say, the value of (classically) liberal institutions like a free press, they would be more widespread. Moreover, philosophical investigation has and will matter for other sciences. The most salient example of this is the impact of philosophy on the Enlightenment, but there are scores of other examples. Note that it is not our view that all philosophical progress leads to these results—certainly not every insight does. However, the historical examples reveal that philosophical work can make a large difference and the expected value of such work can be high in magnitude, even if it may be unlikely that a given insight causes substantive change. In short, when philosophy progresses, it can be expected to be good for philosophy, other fields, and policy.[31]

[30] Although it is rare that policy makers are influenced by philosophers there are notable examples, such as the work of Karl Marx and John Locke. More recently, John Rawls and Peter Singer have had a non-trivial impact.

[31] Our view is especially persuasive if you are a *longtermist* and do not morally discount future persons. See Hilary Greaves & William MacAskill, "The case for longtermism," *GPI Working Paper No. 7-2019*, (2019) for a defense of longtermism.

The progress of philosophy also has extrinsic value in the teaching sphere. Philosophers who teach—which is nearly all of us, it seems—are informed in teaching by what they take to be true. If more of what they take to be true is in fact true, this has a trickledown effect on their students, even if they are incredibly careful to not push their own views on their students. Everything is affected, from the readings they assign to the prominence they give to certain philosophers or theories, the views that they convey to their students as live options, or the objections to views that they take to be most troubling. Even if a given philosopher has a pedagogical outlook where their goal is to develop the intellectual virtues of their students (humility, courage, charitability, and the like), the way in which they develop these virtues will be governed by the truth more often than it would have been, had they not been part of philosophical progress. The proper exercise of all of these virtues is dependent upon one's judgment of the truth, which, if in better touch with the truth itself, is more likely to be informed and thus properly regulative of one's mental habits and tendencies. Having more skilled teachers leads to more philosophical progress which leads to more skilled students and to benefits in countless domains.

There are a variety of principles that license an inference from the significant value of epistemic philosophical progress to the existence of a *prima facie* obligation for philosophers to seek to maximize philosophical progress.[32] This is the sort of duty that is universal in that it applies to all philosophers unless it is overridden by other obligations or made inert by excusing

[32] We must note here that while we have chosen to run our argument in terms of obligation, it could be alternatively formulated without presuming that there are obligations of this kind. For those, like scalar consequentialists, who deny that there are obligations, the argument can be run with a congenial principle that licenses inferences from the significant value of epistemic philosophical progress to the existence of a corresponding degree of rightness. For relevant discussion, see Neil Sinhababu, "Scalar consequentialism the right way," *Philosophical Studies*, 175, (2018): 3131-3144.

conditions.[33] One such principle states that one should seek to maximize the realization of value, whether intrinsic or extrinsic. Although some will undoubtedly find such a principle plausible, and it is certainly sufficient to license this inference, it presumes a rather demanding consequentialist outlook that we cannot defend here. A weaker principle states that one should seek to increase the realization of value, whether intrinsic or extrinsic, so long as the *net* increase in value that would result from one's action is great. Any case where there is a significant gap between the cost one would accrue and the benefits one would realize in acting is a case where one should act. Since we have argued that epistemic philosophical progress has significant value, philosophers who do not face corresponding costs in seeking it should seek it.

That this obligation is a *prima facie* obligation is clear from several considerations. First, some people who could philosophize at a given time nevertheless ought not. We grant that if philosophers should not be philosophizing in the first place, then it does not follow that philosophers should attempt to maximize it. Our position is that *if* the value of philosophizing is sufficient to justify philosophizing at all for a given philosopher at a given time, then they should philosophize such that they seek to maximize philosophical progress, all else equal. In this way, P7 is not a general claim about what should be done or what people should do. The obligation for those who are philosophers to seek to maximize philosophical progress is subject to constraints that, under certain conditions, permit them to refrain from doing philosophy at all. If, for instance, a philosopher ought to take their partner to the hospital, it is not the case that they should seek to maximize philosophical progress at that time.

[33] This sort of obligation is also known as a "component" (and not a "resultant") obligation, as the terms are defined by C.D. Broad ("Some of the Main Problems of Ethics," *Philosophy*, 21, (1946): 99-117), or a "prima facie duty" as W.D. Ross defines it (*The Right and the Good* (Oxford: Oxford University Press, 1930)), as opposed to a "duty proper."

Second, it may be that there is extrinsic value to philosophizing in a way that does *not* attempt to maximize philosophical progress. For instance, there are many benefits of philosophers entering the public sphere and advocating for policies, candidates, and the like. When a philosopher writes an op-ed in a local newspaper arguing for veganism, their audience consists of non-philosophers and yet we can grant that there is significant value to what they are attempting to bring about. Even if no philosophers see their op-ed, the value they are attempting to bring about by writing it might very well outweigh the value they could have attempted to bring about by directing their efforts to help philosophers realize more true or approximately true representations of philosophical propositions.

Although these considerations are reasons for thinking that the obligation for philosophers to seek to maximize philosophical progress is not a categorical one, we maintain that there are relatively few situations where it is *not* the case that philosophers should seek to maximize philosophical progress. That is, we maintain that this obligation is fairly demanding. Most philosophizing is subject to the demands of this obligation.

Philosophical progress has widespread, significant, and positive effects on both research and teaching. To the extent that philosophers maximize philosophical progress, they maximize these salutary effects (and perhaps intrinsic value as well). As a consequence, philosophers have an obligation to attempt to maximize philosophical progress to the extent they can, subject to these countervailing obligations and excuses. In many circumstances where philosophers are philosophizing—such as the classroom, the conference, or the journals—there are not any significant countervailing considerations, as the preceding examples show.

With P7 thus in hand, we can infer C3. From C2 and the preceding claims, it follows that philosophers should theorize about and develop philosophy-specific AI *now*, and they should

utilize it as soon as it is available. (The first conjunct depends on P6, while the second conjunct follows from the fact that in order for philosophically relevant AI to lead to philosophical progress, it must be used.)

This obligation can be understood as one that is capable of being discharged by others. *On the margin*, we believe that there should be more of this kind of work. However, there are also *prima facie* obligations to investigate ethics, epistemology, and the history of philosophy. Clearly, whether these obligations are held by individuals or the philosophical community, they do not entail that every philosopher ought to become an ethicist. Instead, we believe these obligations cease to convert to categorical or all-things-considered obligations when they are sufficiently acted on by others. Hence, although working on developing AI tools should be on a par with investment in traditional philosophical fields, we do not hold that work on these traditional philosophical fields should be radically diminished.

5. Conclusion

In this paper, we have argued that philosophers have an obligation to spend significant effort developing, theorizing about, and using AI to do philosophy. To arrive at this conclusion, we have argued that philosophers have an obligation to maximize epistemic philosophical progress, and that the integration of AI tools with philosophical practice would transform the field and lead to significant progress of this sort.

The path we have taken in this paper involved discussion of specific AI tools that we think would be especially helpful to philosophers, both in general and relative to our epistemic definition of philosophical progress. However, we maintain that there are many alternative paths to the same conclusion that depend on premises concerning different sorts of tools, different definitions of philosophical progress, and, indeed, different goods beyond philosophical progress

entirely. For instance, we hold that it is likely that AI tools developed for teaching philosophy in K-12 educational settings would generate significant value, even if philosophers themselves do not progress in any direct way as a consequence (though the philosophical community likely would). These alternative paths to the same conclusion are a function of the great and multifaceted promise of AI. It should not be underestimated.

Bibliography

Armstrong, Stuart, Sotala, Kaj, & Seán S., Ó hÉigeartaigh. 2014. "The errors, insights and lessons of famous AI predictions – and what they mean for the future." *Journal of Experimental & Theoretical Artificial Intelligence*. 26(3). 317-342.

Amodei, Dario & Danny Hernandez. "AI and Compute." OpenAI. <openai.com/blog/ai-and-compute>. Accessed: October 2, 2020.

Berner, Christopher et al. 2019. "Dota 2 with Large Scale Deep Reinforcement Learning." OpenAI. <https://cdn.openai.com/dota-2.pdf>. Accessed: October 2, 2020.

Bishop, Michael A. & J D Trout. 2005. *Epistemology and the Psychology of Human Judgment*. New York: Oxford University Press.

Bostrom, Nick. 2017. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Broad, C.D. 1946. "Some of the Main Problems of Ethics," *Philosophy*, 21, 99-117.

Brown, Tom B., et al. 2020 "Language Models Are Few-Shot Learners." <https://arxiv.org/abs/2005.14165>.

Buckner, Cameron. 2019. "Deep learning: A philosophical introduction." *Philosophy Compass*. 14 (10). 1-19.

Chalmers, David J. 2015. "Why isn't there more progress in philosophy?," *Philosophy*, 90 (1), 3-31.

Chirimuuta, Mazviita. 2020. "Prediction versus understanding in computationally enhanced neuroscience." *Synthese*.

Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science*. 87:4.

Crisp, Roger & Theron Pummer. 2020. "Effective Justice." *Journal of Moral Philosophy*. 17.

398-415.

Dreyfus, Hubert. 1965. "Alchemy and AI." *RAND Corporation*.

Frances, Bryan. 2017. "Extensive Philosophical Agreement and Progress," *Metaphilosophy*, 48 (1-2), 47-57.

Frances, Bryan & Jonathan Matheson. 2018. "Disagreement." In Zalta, Edward N. (ed.): *Stanford Encyclopedia of Philosophy*, <plato.stanford.edu/entries/disagreement/>. Accessed: October 2, 2020.

Garcez, Artur D'Avila et al. 2019. "Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning." <https://arxiv.org/abs/1905.06088>.

Garcez, Artur D'Avila & Lamb, Luis C. 2020. "Neurosymbolic AI: The 3rd Wave," <https://www.arxiv-vanity.com/papers/2012.05876/>.

Grace, Katja et al. 2018. "When Will AI Exceed Human Performance? Evidence from AI Experts." *Journal of Artificial Intelligence Research.* 62. 729-754.

Greaves, Hilary & William MacAskill. 2019. "The case for strong longtermism." *GPI Working Paper No. 7-2019.*

Hernandez, Danny & Tom Brown. 2020. "Measuring the Algorithmic Efficiency of Neural Networks." <https://arxiv.org/abs/2005.04305>

Jackson, Frank. 2017. "On Connect," in R. Blackford & D. Broderick (eds.): *Philosophy's Future: The Problem of Philosophical Progress*, 51-59.

Jones, Ward E. 2017. "Philosophy, Progress, and Identity," in R. Blackford & D. Broderick (eds.): *Philosophy's Future: The Problem of Philosophical Progress*, 227-239.

Kripke, Saul. 1980. *Naming and Necessity*. Oxford: Basil Blackwell.

Lenat, Douglas B. 1995. "Cyc: A Large-Scale Investment in Knowledge Infrastructure". *Communications of the ACM*, 38 (11).

Lipton, Zachary. 2016. "The Mythos of Model Interpretability". <https://arxiv.org/pdf/1606.03490.pdf>.

MacFarlane, John. 2016. *Relative Truth and its Applications*. Oxford University Press.

Moore, Gordon. 1965. "Cramming More Components onto Integrated Circuits." *Electronics* 114-117.

Morevac, Hans. 1998. "When Will Computer Hardware Match The Human Brain." *Journal of Evolution and Technology* 1.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity.* Hachette Books.

Parfit, Derek. 2011. *On What Matters, vol. 2*. Oxford University Press.

Ricón, José Luis. 2020. "Progress in semiconductors, or Moore's law is not dead yet." *Nintil* <https://nintil.com/progress-semicon/>. Accessed on: October 2, 2020.

Rosenblatt, F. 1958. "The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain." *Psychological Review*. 65(6). 286-408.

Ross, Lewis D. 2020. "How Intellectual Communities Progress." *Episteme*. 1-19.

Ross, W.D. 1930. *The Right and the Good*. Oxford University Press.

Silver, D., Huang, A., Maddison, C. et al. 2016. "Mastering the game of Go with deep neural networks and tree search." *Nature*. 529, 484–489.

Sinhababu, Neil. 2018. "Scalar consequentialism the right way." *Philosophical Studies*. 175. 3131-3144.

Stoljar, Daniel. 2017. *Philosophical Progress: In Defense of Reasonable Optimism*, Oxford

University Press.

Turri, Mark, John, Alfano, and John Greco. 2019. "Virtue Epistemology." In Zalta, Edward N. (ed.), *Stanford Encyclopedia of Philosophy*. <plato.stanford.edu/entries/epistemology-virtue/>. Accessed: October 2, 2020.

Weinberg, Justin. 2020. "Philosophers On GPT-3 (Updated with Replies by GPT-3)." *Daily Nous*, 30 July 2020. <dailynous.com/2020/07/30/philosophers-gpt-3/>. Accessed: October 2, 2020.

Williamson, Timothy. 2008. *The Philosophy of Philosophy*. Wiley Blackwell.

Wilson, Jessica. 2017. "Three Barriers to Philosophical Progress," in R. Blackford & D. Broderick (eds.): *Philosophy's Future: The Problem of Philosophical Progress*, 91-104.

Zhang, Jingqing et al. 2019. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization." <https://arxiv.org/abs/1912.08777>