

# Philosophers Ought to Develop, Theorize About, and Use Philosophically Relevant AI

forthcoming in *Metaphilosophy*  
(penultimate draft; please cite published version)

Graham Clay<sup>1</sup>  
Caleb Ontiveros

*The transformative power of artificial intelligence (AI) is coming to philosophy—the only question is the degree to which philosophers will harness it. In this paper, we argue that the application of AI tools to philosophy could have an 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 a prima facie obligation to spend significant effort in doing so, at least insofar as they should spend effort philosophizing.*

"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."<sup>2</sup>

## 1. Introduction

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. Writing has significantly increased

---

<sup>1</sup> Irish Research Council Government of Ireland Postdoctoral Fellow, School of Philosophy, University College Dublin.

<sup>2</sup> Derek Parfit, *On What Matters* (Oxford: Oxford University Press, 2011), 616.

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.

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 artificial intelligence (AI) would provide a suite of tools that can play a similar role for philosophy. Indeed, the transformative power of AI is coming to philosophy—the only question is the degree to which philosophers will harness it. In this paper, we argue that the application of AI tools to philosophy would likely significantly increase philosophical progress. The role of philosophers in this story is not merely to use AI but also to help develop it and theorize about it.

The structure of this paper is as follows. In section 2, after presenting one viable way of defining philosophical progress, we illustrate how philosophically relevant AI tools would likely lead to significant progress of this sort, if philosophers were to use them. 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 in its development. Finally, in section 4, we argue that philosophers are *prima facie* obligated to spend effort seeking to increase philosophical progress corresponding to the degree that they can increase it via a given method, at least insofar as they should spend effort philosophizing. Given this obligation, philosophers are obligated to spend

significant effort theorizing about and developing philosophically relevant AI *now* and to utilize it as soon as it is available.

## 2. How Artificial Intelligence Would Help Philosophy Progress

Here is the first phase of our argument:

P1. Philosophers who can use philosophically relevant 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, formulate worthwhile questions, and justify philosophical propositions with arguments than philosophers who cannot.

P2. If P1, philosophers who can use philosophically relevant AI tools would be significantly more likely to have more true representations of philosophical propositions—and reach farther intermediate stages on the path to realizing such representations—than philosophers who cannot.

C1. So, philosophers who can use philosophically relevant AI tools would be significantly more likely to have more true representations of philosophical propositions—and reach farther intermediate stages on the path to realizing such representations—than philosophers who cannot.

P3. If philosophers' true representations of philosophical propositions increase over a period—or they reach farther intermediate stages on the path to realizing such representations—then philosophy progresses through that period.

C2. So, it is likely that philosophy would progress significantly more if philosophers can use philosophically relevant AI tools than if they cannot.

There are many different ways to define philosophical progress, but one common way, reflected in P3, 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) of philosophical propositions increase over a period—or if philosophers reach farther intermediate stages on the path to realizing such representations, like those activities listed in P1—then philosophy progresses through that period.<sup>3</sup>

This definition, recently developed and defended by Lewis D. Ross as an ecumenical epistemic definition that allows for partial progress, applies to philosophers' representations of propositions ranging from *eating animal products is wrong* to *there is a distinction between metaphysical grounding and supervenience*. 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.<sup>4</sup> Others focus on philosophers' representations of perennial propositions like God exists.<sup>5</sup> And while some focus on individual philosophers' representations, others argue that it is the epistemic position of the "intellectual community" as a whole that is relevant.<sup>6</sup> On all such views, it is not sufficient for a single isolated philosopher to progress for philosophy to progress, given that the relevant sort of progress is that of the field as a whole.

In this section, we will argue that philosophically relevant 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

---

<sup>3</sup> Lewis D. Ross, "How Intellectual Communities Progress," *Episteme* 18:4, (2021), 740.

<sup>4</sup> For relevant discussion and a defense of a version of this claim, see Bryan Frances, "Extensive Philosophical Agreement and Progress," *Metaphilosophy* 48, (2017), 47-57.

<sup>5</sup> See, e.g., David J. Chalmers, "Why isn't there more progress in philosophy?," *Philosophy* 90:1, (2015), 3-31.

<sup>6</sup> For a discussion of the literature on this point, see Ross, "How Intellectual Communities Progress," 2021.

philosophical progress. Yet, whether or not there are other varieties of progress, epistemic progress of this sort clearly matters in the way that we need for our overall 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 philosophically relevant AI now, and that they should utilize it as soon as it is available. So, in what follows, we will operate with this definition assumed in order to establish our position and reveal a defensible path to our conclusion, but we maintain that philosophically relevant 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 philosophically relevant AI tools*

What exactly is AI? In this subsection, we answer this question and provide some insight into the sorts of AI tools that would transform philosophy, thereby providing the foundation of our justification of P1.

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.<sup>7</sup> 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:

---

<sup>7</sup> A contemporary system exemplifying 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).

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.<sup>8</sup>

Two especially impressive machine learning systems are AlphaGo and GPT-4. In 2015, AlphaGo was the first AI system to beat a professional human player in a full-sized game of Go, beating the European champion 5 games to 0.<sup>9</sup> GPT-4 is a natural language model that is able, with minimal input, to produce paragraphs of content responding to queries about fictional characters, music, politics, philosophy, and many more topics.<sup>10</sup> Its responses are novel but often indistinguishable from those that would be given by a human responding to the same queries.<sup>11</sup>

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.<sup>12</sup>

Presently, our concern is with how such systems could enable philosophers to better complete those tasks listed in P1: locate plausible alternatives to their positions, find literature

---

<sup>8</sup> 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).

<sup>9</sup> Silver, D. et al., "Mastering the game of Go with deep neural networks and tree search," *Nature* 529, (2016), 484–489.

<sup>10</sup> OpenAI, "GPT-4 Technical Report," <<https://arxiv.org/abs/2303.08774>>, (2023). See also the report on the earlier GPT-3: Tom B. Brown, et. al., "Language Models Are Few-Shot Learners," <<https://arxiv.org/abs/2005.14165>>, (2020).

<sup>11</sup> Guo, Biyang et. al., "How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," <<https://arxiv.org/abs/2305.03195>>, (2023).

<sup>12</sup> 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 3<sup>rd</sup> Wave," <<https://www.arxiv-vanity.com/papers/2012.05876/>>, (2020).

relevant to their research, detect errors in their reasoning, generate novel philosophical thoughts, formulate worthwhile questions, and justify philosophical propositions with arguments.

Philosophically relevant AI is AI that can significantly assist philosophers, including in these ways. To be clear, we are not here concerned with the creation of philosophical agents or superintelligence—beings that, in some sense, *do philosophy*—but rather AI tools which, when used by human philosophers, enable them to better complete these tasks.

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 and ideas.
- *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. First, professors could use Recommendation tools that have access to their paper drafts—much like spell-checking or grammar tools do already—in order to find citations and pertinent literature in real time. Second,

Recommendation systems could suggest worthwhile ideas, such as objections, confusions, or questions based on a philosopher's input. In this way, Recommendation systems could enhance philosophers' ability to reason by quickly connecting them to important work and providing questions that may improve philosophical thinking.

*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.<sup>13</sup> 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

---

<sup>13</sup> 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, "Only Connect," in R. Blackford & D. Broderick (eds.): *Philosophy's Future: The Problem of Philosophical Progress* (Oxford: Wiley Blackwell, 2017), 51-59.



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*.<sup>14</sup> Through the use of a Systematizing tool, philosophers would be assisted by AI to filter through potential relations between philosophical positions in order to select and refine relations in useful ways they may not have thought of otherwise. Likewise, philosophers could be directed to potential challenges and counterexamples to their positions by AI tools that identify the consequences of their positions—including those that follow from combining them with others—and the views whose negations are supported or entailed by these consequences. Although Recommendation and Synthesis tools would be powerful, Systematizing tools could be truly transformative to philosophy.

*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. As noted, 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. Imagine being able to talk to (a simulation of) any challenging interlocutor from the philosophical canon!

---

<sup>14</sup> Roger Crisp & Theron Pummer, "Effective Justice," *Journal of Moral Philosophy* 17:4, (2020), 398-415.

*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 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 all philosophically relevant 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 philosophically relevant AI tools on progress*

The foundations for our argument for P1 are found in the preceding subsection. As we have discussed, philosophers who can use the aforementioned philosophically relevant AI tools—from Recommendation to Reasoning tools—would be assisted in every aspect of their work. This assistance would make them more productive and effective regardless of the task they seek to complete because there would be AI tools that directly improve their ability to do 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, formulating worthwhile questions, and justifying philosophical propositions with arguments are among the main tasks of philosophers, and they are tasks that AI tools are 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, philosophically relevant AI tools would make philosophers significantly better at them. In effect, these AI tools would provide an enhanced version of the environment in which philosophers already seek to immerse themselves with graduate school training, 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. Some of these other ways would be conducive to but not constitutive of progress, like making philosophers who are professors more efficient at grading their undergraduates' papers. 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.

P2 links these individual impacts of philosophically relevant AI to a significantly increased likelihood of having more true representations of philosophical propositions—and of reaching farther intermediate stages on the path to realizing such representations. 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 affect their likelihood to have more true representations of philosophical propositions—and to reach farther intermediate stages on the path to realizing such representations. 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, at least in nearly every case. All else equal, philosophers who are better able to do these tasks are significantly more likely to have more true representations of philosophical propositions—and to reach intermediate stages on the path to realizing such 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 representations of philosophical propositions.

Note that P2'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 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, and the epistemic definition of philosophical progress (P3) in hand, it follows that it is likely that philosophy would progress significantly more if philosophers can use philosophically relevant AI tools than if they cannot (C2), whether philosophical progress is

defined in terms of the truth of philosophers' representations of philosophical propositions or their reaching intermediate stages on the path to realizing such representations, whether the relevant propositions are new or perennial, and whether philosophers are treated as a collection of individuals or as an intellectual community.

### 3. The Feasibility of Philosophically Relevant AI

#### *3.1 - The near-term feasibility of philosophically relevant AI*

We have argued that it is likely that philosophically relevant AI *would* help philosophy progress in a significant way. 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:

P4. Within the next decade, it is likely that the technological capabilities of AI will be sufficiently sophisticated to be philosophically relevant.

There is uncertainty about when various sorts of philosophically relevant AI will arrive. However, we believe that the likelihood is more than high enough for our argument to go through. We will give three arguments for this claim about likelihood: the *argument from current progress*, the *argument from compute*, 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.

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 philosophically relevant 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.<sup>15</sup>
- OpenAI's GPT-2 and GPT-3 are released in 2019 and 2020.
- OpenAI releases ChatGPT (based on GPT-3.5) in late 2022 and GPT-4 in 2023.

As an example of how far the GPTs have come, consider GPT-4's response to the query:

"In fewer than 250 words, can you explain how a theist would respond to the problem of evil?":

---

<sup>15</sup> Christopher Berner et al., "Dota 2 with Large Scale Deep Reinforcement Learning," <<https://cdn.openai.com/dota-2.pdf>>, (2019).

The problem of evil is a philosophical argument that questions the existence of an all-powerful and all-good God in light of the existence of evil and suffering in the world. A theist, who believes in the existence of such a God, may respond to this problem in several ways.

Firstly, a theist might argue that evil and suffering are necessary for the greater good or for the development of certain virtues. They may assert that God allows evil to exist in order to give humans free will, as moral choice and responsibility are valuable aspects of human existence.

Secondly, a theist might suggest that evil and suffering are the result of human actions and the consequences of a fallen or imperfect world. They could argue that God created a world with the potential for evil, but it is humanity's misuse of free will that brings about suffering.

Thirdly, a theist may take a perspective that human understanding is limited, and we cannot comprehend the full reasons behind God's actions. They may argue that God's wisdom is beyond human comprehension, and what appears as evil or suffering to us may serve a greater purpose in the divine plan.

Lastly, a theist might emphasize the concept of an afterlife or eternal rewards, where the suffering and injustice experienced in this life can be reconciled and justified. They may argue that God's ultimate justice will be realized in the afterlife, where the righteous are rewarded and the evildoers are punished.

It's important to note that these responses are not exhaustive, and different theists may have varying perspectives on the problem of evil based on their religious beliefs and interpretations.



Despite its imperfections, this technology is already getting to the point where it can be effectively used as a Recommendation and Simulation system.<sup>16</sup> Indeed, the GPTs have progressed substantially since we wrote our first draft of this piece, which included responses from GPT-3 that were significantly less sophisticated and less useful. As the economists Tyler Cowen and Alexander Tabarrok argue in the case of economics:

ChatGPT and related services are extremely useful for a wide variety of services related to learning, or conversely teaching, economics. Think of GPTs not as a database but as a large collection of extremely smart economists, historians, scientists and many others whom you can ask questions. Imagine getting Ken Arrow, Milton Friedman, and Adam Smith in a room and asking them economics questions.<sup>17</sup>

We expect progress on GPT-*n* to continue at a rate that will render the AI-generated examples in this paper to appear rudimentary. And there are other AI-driven classification and summarization tools that are evolving into Synthesis, Formalization, and Systematizing tools.<sup>18</sup> 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 relevant for philosophers 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.<sup>19</sup> Compute power refers to the number of

---

<sup>16</sup> Another example of an effective AI tool for Recommendation is Elicit. See Justin Reppert et. al., "Iterated Decomposition: Improving Science Q&A by Supervising Reasoning Processes," <<https://arxiv.org/abs/2301.01751>>, (2023).

<sup>17</sup> Cowen, Tyler and Tabarrok, Alexander T., "How to Learn and Teach Economics with Large Language Models, Including GPT," *GMU Working Paper in Economics No. 23-18*, (2023).

<sup>18</sup> Jingqing Zhang et al., "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," <<https://arxiv.org/abs/1912.08777>>, (2019).

<sup>19</sup> For an introduction, see Dario Amodei & Danny Hernandez, "AI and Compute," <[openai.com/blog/ai-and-compute](https://openai.com/blog/ai-and-compute)>, 2018.

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.<sup>20</sup> 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.<sup>21</sup> There have been continual and significant algorithmic advances over the last 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).<sup>22</sup> We expect this trend to continue, powering more advanced AI tools like the ones discussed above.<sup>23</sup>

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.<sup>24</sup> (You read that right—*every* task!) Moreover, many researchers are

---

<sup>20</sup> 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.

<sup>21</sup> GPT-3 was trained on V100 GPUs which were introduced in 2017. See Brown et al., "Language Models are Few-Shot Learners."

<sup>22</sup> Hernandez & Brown, "Measuring the Algorithmic Efficiency of Neural Networks."

<sup>23</sup> Suggestively, Hans Morevac 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).

<sup>24</sup> Katja Grace et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," *Journal of Artificial Intelligence Research*, 62, (2018), 729-754.

increasingly worried about the dangers posed by artificial general intelligences (AGIs).<sup>25</sup> 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 (P4). It is worth noting that we are *not* arguing that AI will need to obtain anything like artificial 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 philosophically relevant AI tools. In this subsection, we will motivate the following premise:

P5. Unless philosophers theorize about and develop philosophically relevant AI, it is likely that philosophically relevant AI will be significantly less 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 three ways that philosophers' input would significantly further its development: by assessing its outputs, by shaping its institutional role, and by theorizing about it.

---

<sup>25</sup> 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).

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 would give 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. This is analogous to how OpenAI has developed GPT-*n*.

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.

This is an instance of a broader class of ways in which philosophers' input would further the development of AI tools—namely, by shaping their institutional role. While much of our focus has been on impacts of AI on individual philosophers, there would be significant institutional impacts of AI, too, especially if philosophers were involved in their development. Currently, the prominent AI tools are not integrated within social networks. For instance, as of May 2023, ChatGPT has no social functionality; each user interacts with its interface on their own, and it does not link users to each other in any way. If philosophers were to help develop AI tools, then they could work to develop them in ways that 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. Socially-integrated AI would enable philosophers to communicate better with one another about alternatives to their positions, relevant literature, potential errors in their reasoning, worthwhile questions, 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.<sup>26</sup>

For example, philosophers could help develop AI systems integrated with PhilPapers and PhilPeople so that they are 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 could be designed to facilitate direct communication between users as well, 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.

We maintain that AI tools that enable philosophers to more fruitfully interact on a regular basis with their peers are significantly more useful than those that do not, and they are useful in ways that directly link to our epistemic definition of progress. 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 seem to be tight between increasing philosophers' fruitful interactions, increasing consensus (for the right reasons), and increasing the likelihood of philosophers having true representations—and reaching intermediate stages on the path to realizing such representations. Since it is not possible for both a philosophical proposition

---

<sup>26</sup> 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.

and its negation to be true, if the amount of true representations of philosophical propositions increases over time across all philosophers (or they reach further intermediate stages on the way to realizing such representations), then perhaps will be a "convergence to the truth" amongst philosophers.<sup>27</sup> More importantly, though, it would seem that the relationship 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 representations of philosophical propositions has not significantly increased during it (and they are not likely to have reached intermediate stages on the way to realizing such representations). If, by contrast, such a community of philosophers converges (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.<sup>28</sup> 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 of philosophical propositions than those who are not

---

<sup>27</sup> For discussion, see Chalmers, "Why isn't there more progress in philosophy?". In order for the number of 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 typically be presumed. Clearly, though, there are periods where true representations could increase in virtue of new discoveries but convergence would decrease in virtue of the inertia of old falsehoods.

<sup>28</sup> 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. For a nuanced discussion of ways in which individuals' epistemic improvement can come apart from the improvement of their community (in both directions), see Conor Mayo-Wilson et. al, "The Independence Thesis: When Individual and Social Epistemology Diverge," *Philosophy of Science* 78, (2011), 653-677.

(or are more likely to reach further intermediate stages on the way to realizing such representations). Yet, only if philosophers involve themselves in the development of AI tools would the tools be developed in these useful ways.

The third 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.<sup>29</sup> 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

---

<sup>29</sup> 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).<sup>30</sup> 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. Given rapid progress and the changing technological landscape, we hope that our suggestions appear to be unsophisticated in detail but correct in outline. AI tools can impact philosophy by producing philosophically relevant results, but the extent of their usefulness depends on philosophers.

#### 4. The Obligation to Increase Philosophical Progress with AI Tools

The arguments of the preceding sections established the following claims:

C2. So, it is likely that philosophy would progress significantly more if philosophers can use philosophically relevant AI tools than if they cannot.

P4. Within the next decade, it is likely that the technological capabilities of AI will be sufficiently sophisticated to be philosophically relevant.

P5. Unless philosophers theorize about and develop philosophically relevant AI, it is likely that philosophically relevant AI will be significantly less useful.

In section 2, we established C2, while in section 3 we established P4 and P5. In this section, we will argue for the following moral claim:

---

<sup>30</sup> 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* 199, (2020), 767-790.

C3. Philosophers should spend significant effort to theorize about and develop philosophically relevant AI *now*, and they should utilize it as soon as it is available, at least insofar as they should spend effort philosophizing.

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

P6. Philosophers should spend effort seeking to increase philosophical progress corresponding to the degree that they can increase it via a given method, at least insofar as they should spend effort philosophizing.

As before, here we are relying upon a broadly epistemic notion of philosophical progress.

Philosophy progresses if philosophers have more true representations of philosophical propositions (or reach farther intermediate stages on the way to realizing such representations).

P6 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 as philosophers, or so we will argue.

If philosophers seek to increase 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 increase philosophical progress, they attempt to increase this variety of intrinsic value. Whether they are increasing their own true representations of philosophical propositions or they are increasing those of other philosophers (or reaching

farther intermediate stages on the way), they are increasing the quantity of intrinsically valuable things.

While we are open to the view that philosophical progress has intrinsic value, we are committed to it having significant 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.<sup>31</sup> 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.

---

<sup>31</sup> 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.

Note that it is not our view that all philosophical progress leads to these results—certainly not every insight does. Likewise, though our argument is stated in terms of progress, it is consistent with our view that philosophy has regressed in serious ways and that this has deleterious value for philosophy, science, and culture at large. Specifically, disvalue is created by philosophers promoting falsities instead of truth. 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, policy, and beyond.<sup>32</sup>

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 spend effort seeking to increase it corresponding to the degree that they can increase it via a given method, at least insofar as they should spend effort philosophizing.<sup>33</sup> This is the sort of duty that is universal in that it applies to all philosophers, although, as we will discuss, it can be overridden by other obligations or made inert by excusing conditions that fall under its final clause ("at least insofar as").<sup>34</sup> 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

---

<sup>32</sup> 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.

<sup>33</sup> 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.

<sup>34</sup> 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."

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 and more plausible principle has two parts: first, 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; and second, one should spend effort on such an action corresponding to the degree that it would increase net value. Since we have argued that epistemic philosophical progress has significant value, philosophers who do not face corresponding costs in seeking to increase it should spend effort doing so corresponding to the degree that they can increase it via a given method.

There are several noteworthy features of this obligation. First, we grant that if it is not the case that philosophers should be philosophizing in the first place, then it is not the case that philosophers should be attempting to increase philosophical progress. Our position is that *if* the value of philosophizing is sufficient to justify philosophizing at all, then philosophers should philosophize such that they spend effort seeking to increase philosophical progress corresponding to the degree that they can increase it via a given method. In this way, P6 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 increase 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 increase philosophical progress at that time. Hence our qualifier "at least insofar as they should spend effort philosophizing."

Second, and relatedly, it may be that there is extrinsic value to philosophers promoting philosophy in a way that does *not* attempt to increase 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 progress. As we are using the term, this is not "philosophizing," and any changes it brings in non-philosophers is not philosophical progress.

Third, this obligation is a burden of the community of philosophers, so it should 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 parallel obligations to investigate ethics, epistemology, and the history of philosophy. Clearly, they do not entail that every philosopher ought to become an ethicist; they cease to be obligations for the community when they are sufficiently acted on by others. (Hence, although working on and using 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.)

With P6 thus in hand, we can infer C3. From C2 and the preceding claims, it follows that philosophers should spend significant effort to theorize about and develop philosophically relevant AI *now*, and they should utilize it as soon as it is available, at least insofar as they should spend effort philosophizing. (The first conjunct depends on P5, while the second conjunct follows from the fact that in order for philosophically relevant AI to lead to philosophical progress, it must be used.) Since the likely philosophical relevance of AI tools is so significant—and so uniquely significant—the demand on philosophers is accordingly high.

## 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, at least insofar as they should spend effort philosophizing. To arrive at this conclusion, we have argued that philosophers have an obligation to increase epistemic philosophical progress corresponding to the degree that they can increase it via a given method, 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. These alternative paths to the same conclusion are a function of the great and multifaceted promise of AI. It should not be underestimated.<sup>35</sup>

---

<sup>35</sup> For feedback on earlier drafts of this paper, we would like to thank Daniel Immerman, as well as an anonymous reviewer at and the editors of *Metaphilosophy*. Some of the research conducted in this publication was funded by the Irish Research Council under grant number GOIPD/2021/413.

### Bibliography

- Armstrong, Stuart, Kaj Sotala & 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.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* 199. 767-790.
- Cowen, Tyler & Alexander T. Tabarrok. 2023. "How to Learn and Teach Economics with Large Language Models, Including GPT." *GMU Working Paper in Economics No. 23-18*.
- Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of*



*Science* 87:4. 568-589.

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. 47-57.

Frances, Bryan & Jonathan Matheson. 2018. "Disagreement." In Zalta, Edward N. (ed.): *Stanford Encyclopedia of Philosophy*, <[plato.stanford.edu/entries/disagreement/](http://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 & Luis C. Lamb. 2020. "Neurosymbolic AI: The 3<sup>rd</sup> 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*.

Guo, Biyang et. al. 2023. "How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," <<https://arxiv.org/abs/2305.03195>>.

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

Jackson, Frank. 2017. "Only 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.
- Mayo-Wilson, Conor, Kevin J.S. Zollman & David Danks. 2011. "The Independence Thesis: When Individual and Social Epistemology Diverge." *Philosophy of Science* 78. 653-677.
- 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.
- OpenAI. 2023. "GPT-4 Technical Report," <<https://arxiv.org/abs/2303.08774>>.
- 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.
- Reppert, Justin et. al. 2023. "Iterated Decomposition: Improving Science Q&A by Supervising Reasoning Processes," <<https://arxiv.org/abs/2301.01751>>.
- 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. 2021. "How Intellectual Communities Progress." *Episteme* 18:4. 738–756.
- Ross, W.D. 1930. *The Right and the Good*. Oxford University Press.
- Silver, David 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, John, Mark Alfano & John Greco. 2019. "Virtue Epistemology." In Zalta, Edward N. (ed.), *Stanford Encyclopedia of Philosophy*. <[plato.stanford.edu/entries/epistemology-virtue/](https://plato.stanford.edu/entries/epistemology-virtue/)>. 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>>