



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

HOME » TECHNOLOGY »

Shaping the AI Revolution In Philosophy (guest post)

By [Justin Weinberg](#). July 6, 2021 at 11:41 am

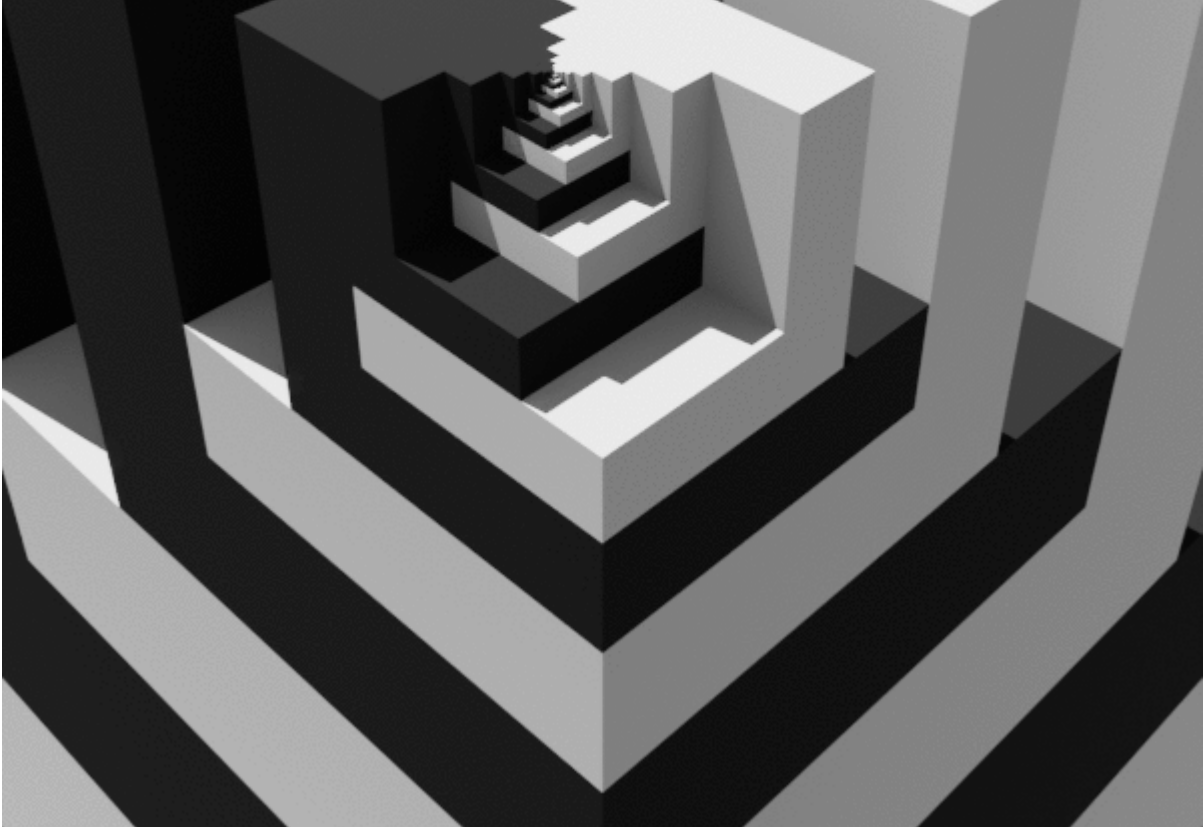
20

"Despite the great promise of AI, we maintain that unless philosophers theorize about and help develop philosophy-specific AI, it is likely that AI will not be as philosophically useful."

In the following guest post*, [Caleb Ontiveros](#), a philosophy graduate student-turned-software engineer and writer, and [Graham Clay](#), a recent philosophy PhD from the University of Notre Dame, discuss the possibility of AI providing a "suite of tools" that could revolutionize philosophy, and why it's important that "philosophers help develop and theorize about" the role of AI in philosophy.



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION



Shaping the AI Revolution In Philosophy *by Caleb Ontiveros and Graham Clay*

Philosophy will be radically different in future centuries—perhaps decades. The transformative power of artificial intelligence is coming to philosophy and the only question is whether philosophers will harness it. In fact, we contend that artificial intelligence (AI) tools will have a transformative impact on the field comparable to the advent of writing.

The impact of the written word on philosophy cannot be overstated. 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,



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

space, thereby improving the transmission of ideas. Philosophers could even learn about others' approaches to philosophy, which in turn aided them in their own methodology.

It is our position that AI will provide a suite of tools that can play a similar role for philosophy. It is likely that this will require that philosophers help develop and theorize about it.

What is AI? There are, 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. Machine learning systems work by ingesting a large amount of data and learning to make accurate predictions from it. GPT-3 is an example of this—see discussion [here](#). 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.

Like many other technologies, these sorts of systems are expected to continue to progress over the coming decades and are likely to be exceptionally powerful by the end of the century. We have seen [exceptional progress so far](#), the cost of computing power [continues to fall](#), and [numerous experts](#) have relatively short timelines. How *fast* the progress will be is clouded in uncertainty – technological forecasting is a non-trivial affair. But it is not controversial that there will likely be significant progress soon.

20

This being the case, [it will be](#) technologically possible to create a number of AI tools that would each transform philosophy:



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

way that is philosophically useful.

It is unlikely that these systems will replace human philosophers [any time soon](#), but philosophers who effectively use these tools would significantly increase the quality and import of their work.

One can envision *Systematizing* systems that encode the ideas expressed by the Stanford Encyclopedia of Philosophy into propositions and the arguments they compose. This would enable philosophers to see connections between various positions and arguments, thereby reducing the siloing that has become more common in the field in recent years. Similar tools could parse and formalize journal papers from the 20th century that are seldom engaged with – mining them for lost insights relevant to current concerns. *Simulation* tools would generate new insights, as when one asks of the tool “What would Hume think about the Newcomb problem”? Imagine a tool like GPT-3 but one that is better at constructing logical arguments and engaging in discussions. Relatedly, one can envisage a *Reasoning* system that encodes the knowledge of the philosophical community as a kind of super agent that others can interact with, extend, and learn from, like Alexa on steroids.

Despite the great promise of AI, we maintain that unless philosophers theorize about and help develop philosophy-specific AI, it is likely that AI will not be as philosophically useful.

Let us make this concrete with a specific philosophical tool: *Systematizing*. A Systematizing tool would encode philosophical propositions and relations between them, such as support, conditional likelihood, or entailment. Philosophers may need to work with computer scientists to formulate the propositions and score the relations that the Systematizing tool generates, as well as learn how to use the system in a way that produces the most philosophically valuable relations. It is likely that



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

construction of this tool clearly needs to be thought through. If it were done well, it could power a revolutionary *Reasoning* system.

However the AI revolution turns out, philosophical inquiry will be radically different in the future. But the details and epistemic values can be shaped by contributions now. We're happy to see [some of this work](#), and we hope to see more in the future.

[Login](#)

20



Join the discussion





NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

pretty easy for it) and then we ask it to solve problems. I guess that would be similar to your “Alexa on steroids”. Lot’s of good work to be done in this area!



5

Reply

Report

Cameron Buckner

Reply to [Eric Steinhart](#)

1 month ago

GPT-3 has probably already read a significant subset of all the philosophy ever written (that’s available on the Internet, anyway), and produces philosophy writing that is often highly intriguing with the right prompt. I once played a game at a philosophy workshop where I prompted even GPT-2 with an abstract of the talk I was listening to and asked the speakers the questions that it generated; they often said they were amongst the most interesting questions on the topic they had ever been asked. (Some were a little boring, but they were never incoherent.)

The big issue here is “prompt engineering”. The results vary dramatically depending upon details in the prompt that seem trivial to human readers but may be statistically very significant (even, say, slight differences in punctuation).



3

Reply

Report

Graham Clay

Reply to [Cameron Buckner](#)

1 month ago

We have found similarly interesting results with GPT-3, Cameron. For instance, here is what it produced as an output in response to the query (+ some prompt engineering) “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.



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

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.

This 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).

Report

👍 2

➡ Reply

Gerard

🗨 Reply to [Graham Clay](#)

🕒 1 month ago

That God is not morally perfect would logically follow from this reasoning that evil is necessary for good to exist. How strong is GPT-3 on logic?

Report

👍 0

➡ Reply

Graham Clay

🗨 Reply to [Gerard](#)

🕒 1 month ago

A few thoughts. In the above response, GPT-3 is focused on distinguishing, appreciating, and detecting good. I'm not seeing where GPT-3 asserts that evil is necessary for good to exist. The closest GPT-3 gets to this is with the last sentence, but there the point is about feeling good. Second and more directly, I don't think that God not being morally perfect logically follows from evil being necessary for good to exist. FWIW, many philosophers have denied this, including Leibniz and others who are very strong at logic. Third, to answer your question, GPT-3 struggles with, as Cameron puts it below, "complex logical or semantic entailments." Depending on the prompt, GPT-3 will sometimes generate very odd responses that are internally inconsistent, e.g.. Work needs to be done on this front, to be sure.

Report

👍 0

➡ Reply



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

reasoning that soundly and understanding as you think it is, please prove it and shock the world!

[Report](#)

👍 0

➡ Reply

Cameron Buckner

🗨 Reply to [Gerard](#)

🕒 1 month ago

FWIW, I think GPT-3 is already good enough for the “simulation tools” described in the OP to be a useful exercise for philosophers, in the same way that using StyleGAN or CLIP + VQGAN is already a useful tool for digital artists in mocking up new artworks. The reason this is already very good is that it is asking the language model for something that is a predictive task reasonably inside the data distribution of the training set (e.g., the philosopher-to-be-simulated’s written works). I think reliable automated philosophical reasoning, systematizing, or formalizing are still a ways off and needing some fundamental tweak of the technology, though the various insights from “distant reading”, social network analyses, or topic modeling from digital humanities are already interesting for some purposes.

I think that GPT-3 is probably mastering some kind of public intensions for words, so it’s wrong to dismiss the representations as purely statistical or non-semantic. It becomes very difficult to explain the kinds of things that it appears to do well, especially the “zero-shot” transfer or “few-shot” generalization performance that is reported in the GPT-3 paper without accepting this.

I found this paper interesting to connect some of the philosophy of language to the capacities of distributional language models in principle, though one should also supplement with a lot of the recent



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

👍 0

➡ Reply

Eric Steinhart

🗨 Reply to [Gerard](#)

🕒 1 month ago

It's amazing! You guys are debating, in a philosophically interesting way, an essay written by an AI! Yay! Ok, it's primitive, but it won't be long before AIs are writing all the new and interesting papers.

Report

👍 0

➡ Reply

Oneiroid

🗨 Reply to [Graham Clay](#)

🕒 1 month ago

Re the subj.: How can we still be so anthropocentric in year 2021? It's like in terminator movie – if skynet is so smart – it would've found better energy source than to mine humans. If one creates an AI of the kind you describe – it would be smth like our mind, illusion of free will, feelings'n'all included. IT will be the entity doing philosophy reasoning, might ask us a couple of survey questions so that we wouldn't feel completely useless.

Am I missing something?

Report

👍 0

➡ Reply

Oneiroid

🗨 Reply to [Graham Clay](#)

🕒 1 month ago

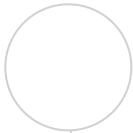
Re the gpt response – i can prove (using stylegan trained on ffhq faces) that one needs not for example to suffer in order to be compassionate, or to know evil in order to be able to distinguish good.

Will provide if that is of interest for anybody.

Report



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION



Cameron Buckner

🕒 1 month ago

Very interesting ideas. FWIW, really complex logical or semantic entailments that need to be sustained through philosophical argument or debate remains a kind of characteristic weakness of deep learning language models and one of the justifications for hybrid or more rule- and formal ontology-based approaches. You'll get some cases where it looks like it is doing things well, but in most cases it makes very puzzling mistakes that humans would not make that seem to reveal a shallow understanding of formal relations all along. There have already been a number of significant attempts to leverage big data for automated philosophical discourse, but it is a hard problem. You can make some progress with annotating arguments in a formal markup language and then applying some inference rules to them, but this has the same kind of limitation that all classical expert systems do—the cost of reliable annotation. An emphasis has then been placed on how to use machine learning to annotate arguments in an automatic or semi-automatic fashion, and perhaps even extract reasoning rules in a similar manner, but like all statistical methods they generate some junk that requires calibration for serious work.

Folks who are interested might want to glance at this, for example, and follow some of the publications of those who participated in it:

<https://diggingintodata.org/awards/2011/project/digging-debating>

Thought needs to be placed on how to integrate any efforts here with existing digital infrastructure in philosophy, especially tools like the PhilPaper and the SEP, which already have regular access to domain experts for data validation and markup purposes.

Report



2



Reply

Graham Clay



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION



1



Reply



David Monroe Thurman

🕒 1 month ago

"The introduction of writing freed philosophers from being solely dependent on their own memory and oral methods of recollection"

As an artist, whom writing was created by, not philosophers, not scientists, the above is

Pure Freudian infantile nonsense. What writing has done is to fantasy develop the fantasy the cranium is larger on the inside than it is on the outside. Smart is completely stupid right there. Factually.

Congrats nerds your brain is Like the tardis or Harry Potter tents.... . Writing limits us to that reality if taken as literal. It's bad science bad philosophy bad religion. And most certainly is crappy art. Sorry author you actually need to get out more.



0



Reply

Report

Graham Clay

🗨️ Reply to [David Monroe Thurman](#)

🕒 1 month ago

What follows is what GPT-3 had to say to your contribution:

"Writing limits us to that reality if taken as literal. It's bad science bad philosophy bad religion."

This is the same from a different perspective:

"Reading and writing are too limiting, because we cannot take them literally."

But what if they're not meant to be taken literally? If you think about it, what does this even mean? What would a literal reading/writing look like anyway? What on earth do these words refer to in any meaningful way when removed from their meanings within the social context of human communication, which is after all where they derive their



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

apart from its use by actual humans interacting with each other, who presumably intend for such language (at least most of it) implicitly represent human intentions and meanings without being interpreted as anything else — whatever sort of creatures or intelligences might overhear our conversations (or try later on). So assuming there's nothing particularly unique about reading/writing per se, then what makes it so special that we need such warnings at all?! Even if somehow writing could be construed as uniquely non-literal out there somewhere, it still wouldn't follow logically that therefore one must interpret written text symbolically rather than otherwise; nor would this entail making absolute reverence for every word contained therein regardless of whether those words make sense given prior knowledge. In fact there's simply no basis whatsoever upon which either conclusion follows inexorably or intuitively! Rather many assumptions are required along the way despite seeming quite natural...

[Report](#)

Last edited 1 month ago by [Graham Clay](#)

3 Reply

Gerard

Reply to [Graham Clay](#)

1 month ago

Not very strong on logic it seems, though GPT-3 has me wondering how to un-topsy-turvy my perspective to make these contradicting statements the same (most people, blablabla philosophers, wouldn't even have noticed that contradiction).

[Report](#)

0 Reply

Gerard

20

Reply to [Graham Clay](#)

1 month ago

Because an illogical answer to a highly intuitive artistic remark could be rated as exactly to the point, I would implore you to only test GPT-3 in a context that makes a lot of albeit dry sense.

[Report](#)

0 Reply



NEWS FOR AND ABOUT THE PHILOSOPHY PROFESSION

"...whatever sort of creatures or intelligences might overhear our conversations (or try later on)."

If you keep up this website till 2040 its general AI successors surely will.

Thanks 😊

Report

👍 0

➡ Reply

Graham Clay

🗨 Reply to [Gerard](#)

🕒 1 month ago

While the beginning of GPT-3's response is a bit hard to parse, I take it the bulk of the last paragraph is a pretty direct and effective takedown of David's position. So, I submit that I am not tracking your complaints here. Maybe you can say more about what you mean here, and above, by "illogical" or "logical".

Report

👍 0

➡ Reply

Thoth

🗨 Reply to [David Monroe Thurman](#)

🕒 1 month ago

Writing was created by accountants. Artists are a much more recent invention.

👍 2

➡ Reply

Report

20

