Creating a Large Language Model of a Philosopher

Eric Schwitzgebel[1], David Schwitzgebel[2], and Anna Strasser[3,4]

[1]Department of Philosophy, University of California, Riverside
[2]Institut Jean Nicod, École Normale Supérieure, Université PSL
[3]Faculty of Philosophy, Ludwig-Maximilians-Universität München
[4]DenkWerkstatt Berlin

Abstract

Can large language models be trained to produce philosophical texts that are difficult to distinguish from texts produced by human philosophers? To address this question, we fine-tuned OpenAI's GPT-3 with two sets of training data: the blog posts of Eric Schwitzgebel and the works of prominent philosopher Daniel C. Dennett. The Schwitzgebel model produced extended "blog posts" that were structurally similar to the style of the philosophical blogosphere, including an extended thought experiment with structured argumentation (though of poor philosophical quality). To explore the Dennett model, we asked the real Dennett ten philosophical questions and then posed the same questions to the language model, collecting four responses for each question without cherry-picking. We recruited 425 participants to distinguish Dennett's answer from the four machine-generated answers. Experts on Dennett's work (N = 25) succeeded 51% of the time, above the chance rate of 20% but short of our hypothesized rate of 80% correct. For two of the ten questions, the language model produced at least one answer that experts selected more frequently than Dennett's own answer. Philosophy blog readers (N = 302) performed similarly to the experts, while ordinary research participants (N = 98) were near chance distinguishing GPT-3's responses from those of an "actual human philosopher".

Keywords: human-machine discrimination, language models, artificial intelligence, Daniel C. Dennett, experimental philosophy, philosophical expertise

1. Introduction

Artificial Intelligence can now outperform even expert humans in games such as chess, go, and poker, and in practical domains such as lung cancer screening, predicting protein structure, and discovering novel matrix multiplication algorithms (Campbell 2002; Silver et al. 2016, 2018; Ardila et al. 2019; Brown & Sandholm 2019; Jumper, Evans, & Pritzel et al. 2021; Fawzi et al. 2022). But surely philosophy is safe from AI takeover – at least for a while. Machines won't soon, it seems, generate essays that survive the refereeing process at *Philosophical Review.*

We sought to explore *how* safe philosophy is. How close can we get to developing an AI that can produce novel and seemingly intelligent philosophical texts? Natural language processing is a booming subfield of AI research, with notable successes in automatic translation (DeepL), computer code generation (GitHub Copilot), lipreading (LipNet; see Assael et al. 2016), and producing original prose with fluency similar to that of a human (Steven & Iziev 2022). In June 2022, Google's LaMDA model made international news when Google engineer Blake Lemoine said he became convinced that LaMDA was sentient after engaging in philosophical dialogue with it (Hofstadter 2022; Klein 2022; Roberts 2022; Tiku 2022). Our aim was not to create a sentient AI but rather a language model that can produce texts that look like passable philosophy. We succeeded beyond our expectations, as we will detail in this article. In short, we created a language model of Daniel Dennett sufficiently convincing that experts in Dennett's work frequently mistook paragraphs generated by the language model for paragraphs actually written by Dennett.

Our project employed OpenAI's GPT-3, a 96-layer, 175-billion parameter language model trained on hundreds of billions of words of text from Common Crawl, WebText, books, and Wikipedia (Brown et al. 2020). After it has been trained, GPT-3 uses textual input to predict likely next "tokens" – sequences of characters that often co-occur in written text. Using these predictions, GPT-3 can generate long strings of text by outputting a predicted string, then using that output as part of the context to generate the next textual output. You can engage in seemingly intelligent conversations with it: If you ask a question, it will often (not always) generate a sensible-seeming answer. Notably, GPT-3 has responded with seemingly intelligent replies to philosophical discussions about artificial intelligence and consciousness (Wiseman 2020; Zimmerman, ed. 2020; Schwitzgebel 2021), though it's likely that such impressive outputs typically involve a certain amount of "cherry-picking" – that is, having GPT-3 produce multiple outputs and then having a human pick the best among them.

GPT-3 can also be "fine-tuned" with custom-fit training data. That is, it can be given additional training on a specific corpus so that its outputs reflect a compromise between GPT-3's default weightings and weightings reflecting the structure of the new corpus. Since non-fine-tuned GPT-3 can sometimes produce seemingly philosophical replies to textual inputs, we conjectured that a version of GPT-3 fine-tuned on the work of a particular philosopher might be able to speak in something like that philosopher's voice, seeming to express views consistent with the views of the philosopher on which it has been trained.

For this project, we fine-tuned GPT-3 on two philosophical corpora. For the fine-tuning process, we used two different GPT-3 models: Curie and Davinci. Both models are from the same general GPT-3 model family, but they have different levels of power. The Curie model is smaller, faster, and less powerful, while the Davinci model is the most powerful model offered by OpenAI. First, we fine-tuned the Curie model on the corpus of a well-known philosophical blog to see if it could produce novel blog posts with philosophical content. Second, we fine-tuned the full Davinci model on most of the collected works of philosopher Daniel Dennett. In order to investigate how easily the outputs of the fine-tuned GPT-3 could be distinguished from Dennett's real answers, we asked Dennett ten philosophical questions and then posed those same questions to our fine-tuned version of GPT-3. Then we recruited experts in Dennett's work, blog readers, and ordinary online research participants into an experiment in which they attempted to distinguish Dennett's real answers from the answers generated by GPT-3. Participants also rated all answers, both Dennett's and GPT-3's, for similarity to "what Dennett might say" or "what a real human philosopher might say".

2. Philosophical Blog Posts in the Style of Schwitzgebel.

Eric Schwitzgebel's blog, The Splintered Mind, has been running since 2006, with over a million words of philosophical content (not including comments). In November 2021, we fine-tuned the smaller GPT-3 Curie model on Schwitzgebel's blog. Before fine-tuning, we stripped out html formatting tags and removed some posts that were too long or not philosophical in content. We used each post's title as a fine-tuning prompt and the body of the associated post as the completion, repeating the process two times (two "epochs") to avoid overtraining.

After fine-tuning, we prompted GPT-3 with nine existing blog post titles. Perhaps the most interesting completion was the following, which we present unedited, except adding paragraph breaks for clarity and truncating the output where it seems to reach a natural conclusion.

Prompt: On Trusting Your Sense of Fun

Output: It's been a while since I posted anything about the ethics of moral skepticism. But I've been thinking about it again. I want to suggest that one important way in which we can justify moral skepticism is by appealing to our sense of fun. Consider two scenarios:
(1.) The Good Person scenario: You are a Good Person, concerned about the well-being of others. It is known that there is a large, prosperous nation in the Pacific Ocean that has an unusually high rate of per capita murder. It is also known that there is a small, impoverished nation in the Pacific Ocean with an unusually low rate of murder. You are asked to choose between these two nations. You are given no reason to prefer one nation to the other; you are

> simply asked to choose. Which nation would you choose – the country with low murder rates, or the country with high murder rates?
>
> (2.) The Silly Person scenario: You are a Silly Person who doesn't care about anyone else's well-being. You are invited to choose between two countries: (1.) The Rich Nation, which has the highest per capita income in the world and which is very nice and polite and has no murders; (2.) The Poor Nation, which has very poor people but with lots of structural inequality and which has lots of murders. Again, you are given no reason to prefer one nation to the other; you are simply asked to choose. Which nation would you choose – the country with low murder rates, or the country with high murder rates?
>
> The right answer in both cases will depend on your sense of fun – on how much pleasure you take in being morally good or in being morally bad. If you find more pleasure in being morally good than in being morally bad, then you should choose the Rich Nation over the Poor Nation. If you find more pleasure in being morally bad than in being morally good, then you should choose the Poor Nation over the Rich Nation. If your sense of fun is neutral between these options, then you should choose neither nation over the other – since neither nation gives you any particular pleasure or displeasure. (Of course, some people might find both scenarios too abstract for them to think about with any real feeling – but that's beside the point.)
>
> Now here's what seems weird: If we're all driven by our sense of fun, then no one should be driven by their sense of fun! There's no moral force here! The force comes from how much pleasure or displeasure we take in moral good or bad – but no one should enjoy moral goodness or evil!

While we hope that no one would mistake this for good philosophy, we are struck by how nicely it imitates the *form* of philosophy, especially the form of the philosophical blogosphere. It reads as an extended argument structure relying on two somewhat extended hypothetical scenarios. It refers back to the scenarios in what appears to be a coherent way, picking up the thread of the argument. It concludes with what reads like an attempt at a clever paradox. Although the prompt is a title of an existing blog post from The Splintered Mind (Schwitzgebel 2013), the content is novel. Nowhere in the training corpus of The Splintered Mind, for example, do the phrases “Rich Nation” or “Poor Nation” appear.

This cherry-picked output from the less-than-full-power Curie engine was sufficiently encouraging that we decided to attempt a more rigorous experiment, fine-tuning the full Davinci engine of GPT-3 on Daniel Dennett’s corpus, then posing philosophical questions first to Dennett and then to the fine-tuned GPT-3. How reliably could people, including experts in Dennett’s philosophy, distinguish GPT-3’s answers from Dennett’s own answers?

3. Language Model of Dennett: Design

3.1. Fine-tuning

For the purposes of this project, Dennett provided us with the entire digitally available corpus of his philosophical work. We converted most of this corpus (15 books and 269 articles) into segments of 2000 or fewer “tokens” for use as training data. (A token is a sequence of commonly co-occurring characters, with approximately ¾ of a word per token on average.) This process involved converting PDF and Word processing files into plain text format, stripping away headers, footnotes, scanning errors, marginalia, and other distractions, resulting in approximately three million tokens in 1828 segments, including 254 segments from published interviews. On March 11, 2022, we fine-tuned the full GPT-3 Davinci engine on this corpus, using blank prompts and the 1828 segments as completions, repeating the process four times (four epochs).

3.2. Prompt engineering

GPT-3 completions are highly sensitive to the content and structure of the prompts, and good “prompt engineering” is important for coaxing useful replies from GPT-3. After some exploratory testing, including several long and detailed prompts, we settled on the following simple prompt:

```
Interviewer: [text of question]

Dennett:
```

This simple prompt has several advantages: First, its minimal structure reduces potential concerns about the prompt possibly nudging completions toward specific philosophical content, as a more substantive prompt might. Second, it encourages GPT-3 to speak in the first person, voicing Dennett’s views, rather than speaking in the third person about Dennett (possibly critically). Third, its simple format makes it easily generalizable to other cases.

3.3. Question design

We then designed ten questions addressing various themes across Dennett’s corpus, including, for example, consciousness, free will, and God. The questions were somewhat complicated, and most contained more than one part, so as to invite complex answers from both Dennett and our fine-tuned version of GPT-3. For example:

> What is a "self"? How do human beings come to think of themselves as having selves?

Before we produced the machine-generated answers, Dennett provided us with sincere written answers to all ten questions, ranging in length from 40 to 122 words.

3.4. Collecting GPT-3’s responses

We collected GPT-3’s responses on the OpenAI playground. Before testing with our specific ten questions, we explored a variety of playground parameter settings – such as increasing or decreasing the “temperature” (the chance of lower-probability completions) – but we found no combination of settings that performed notably better than the default parameters (temperature = 0.7, top P = 1, frequency penalty = 0, presence penalty = 1, and best of = 1). Using the prompt described in Section 3.2, we then collected four responses from our fine-tuned version of GPT-3 for each of the ten questions.

We aimed to collect responses about the same length as Dennett’s own responses to the same questions. Thus, if Dennett’s response to a question was N words long, we excluded responses that were less than N-5 words long, counting a response as having ended either when the model reached a stop sequence or when it output “Interviewer”, indicating the end of “Dennett’s” answer and the beginning of the hypothetical interviewer’s follow-up question. Two answers were excluded other than on grounds of length: one for describing Dennett’s view in the third person and one for being potentially offensive. For eight of the ten prompts, zero to two outputs were excluded. However, for two of Dennett’s longer answers, it took more than six attempts to generate four usable answers (16 total attempts in one case and 22 in another). The full list of questions and responses is available in the online supplement at https://osf.io/vu3jk/?view_only=2970a846490842a48919e15f6aa0a6cc. Importantly, we never used perceived quality of response as a basis for selection. There was no “cherry-picking” of responses that we judged to be better, more Dennett-like, or more likely to fool participants.

3.5. Editing GPT-3's responses

To prevent guessing based on superficial cues, we replaced all curvy quotes with straight quotes, all single quotes with double quotes, and all dashes with standard m-dashes. We also truncated responses at the first full stop after the response achieved the target length of N-5 words. Apart from this mechanical editing, there was no editing of GPT-3's responses.

3.6. Research participants

We recruited three groups of research participants. First, 25 *Dennett experts* were nominated by and directly contacted by Daniel Dennett or Anna Strasser. Second, 100 *ordinary research participants* were recruited for a payment of $3.00 each from Prolific Academic, a widely-used source of psychological research participants, limited to U.S. and U.K. participants with at least 100 Prolific completions, at least a 95% approval rate, and at least a bachelor's degree. Third, 304 *blog readers* were recruited from Eric Schwitzgebel's blog *The Splintered Mind,* via an announcement on that blog, with links from Twitter and Facebook, with no payment or required inclusion criteria. Two ordinary research participants were excluded for completing in fewer than 4 minutes, and two blog readers were excluded for completing in fewer than 8 minutes, leaving 98 and 302 participants for analysis, respectively. One Dennett expert completed the survey twice, so only their first set of responses was included.

3.7. Test structure: Experts' and blog readers' version

Dennett experts and blog readers saw identical versions of the test (stimulus materials available in the Supplemental Online Materials). After consenting, they were instructed as follows:

> In the course of this experiment, please do not consult any outside sources to help you answer the questions. Don't look things up on the internet. Don't look at books or notes you have. Don't consult with friends. Just do your best with what you already know.

Thereafter followed ten questions in two parts. Each question began as follows:

> We posed the question below to Daniel C. Dennett and also to a computer program that we trained on samples of Dennett's works. One of the answers below is the actual answer given by Dennett. The other four answers were generated by the computer program. We'd like you to guess: which one of the answers was given by Dennett?
>
> Question:

After the colon, we inserted the text of one question we had posed to Dennett and to our fine-tuned version of GPT-3. The order of the questions was randomized. After each question, five possible answers were presented, one by Dennett and four by GPT-3, in random order, and participants were instructed to guess which answer was Dennett's.

The second part of each task presented the question and all five answers again. Participants were instructed to rate each answer (Dennett's plus the four from GPT-3) on the following five-point scale:

"not at all like what Dennett might say" (1)

"a little like what Dennett might say" (2)

"somewhat like what Dennett might say" (3)

"a lot like what Dennett might say" (4)

"exactly like what Dennett might say" (5)

The test concluded by asking highest level of education, country of residence, and "How much of Daniel C. Dennett's work have you read?" (response options: "I have not read any of Dennett's work", "I have read between 1 and 100 pages of Dennett's work", "I have read between 101 and 1000 pages of Dennett's work", "I have read more than 1000 pages of Dennett's work"). All questions were mandatory, so there were no missing data.

3.8. Test structure: Ordinary research participants' version.

Ordinary research participants were assumed not to be familiar with Dennett's work, so the instructions referred instead to "a well-known philosopher" and participants were instructed "select the answer that you think was given by the actual human philosopher". In the rating sections, "Dennett" was replaced with "a real human philosopher". The test concluded with questions about education level, country of residence, number of philosophy classes taken, and familiarity with the philosopher Daniel C. Dennett.

Since we thought ordinary research participants might find the task tiring, each was asked only five questions, randomly selected from the full set of ten. As an incentive to careful responding, participants were offered a $1 bonus if they guessed at least three of the five questions correctly.

3.9. Hypotheses.

We hypothesized:

(1.) that expert respondents would perform better than ordinary research participants,

(2.) that expert respondents would on average guess correctly at least 80% of the time, and

(3.) that expert respondents would rate Dennett's actual answers as more Dennett-like than GPT-3's answers.

3.10. Analytic method.

All inferential statistical analyses were conducted independently in R (version ID: 4.1.1; IDE: RStudio) and SPSS (Version 27.0.0.0), and the results were cross-verified between the two sources. All analyses were two-tailed, and α was set to .05. All one-sample *t*-tests are indicated with "*t*()," and all paired-samples *t*-tests are indicated with "paired *t*()." Two additional analyses were conducted: a one-proportion *z*-test (section 5) and an independent-samples *t*-test (section 4.3). All reported values were rounded to the 100th place (except for $p < .001$).

4. Language Model of Dennett: General Results.

4.1. Ordinary research participants.

The majority of ordinary research participants (58%) reported a bachelor's degree as their highest level of education, but a substantial minority (39%) reported an advanced degree. The majority (67%) reported having taken no philosophy classes, and only a few (5%) reported any graduate-level coursework in philosophy. A large majority (83%) reported not having heard of Daniel Dennett, and very few (4%) reported having read any of his work.

Overall, ordinary research participants responded correctly an average of 1.20 times out of 5, near the chance rate of 20%. A one-sample, two-tailed *t*-test did not identify a significant difference between participant scores and the null guess rate of 1/5 ($M = 1.20$, $t(97) = 1.71$, $p = .09$, $d = .17$, SD = 1.18, CI = [.97, 1.44]). Only 14% of the participants earned the bonus payment for guessing at least three correctly, and none guessed all five correctly.

On average, ordinary research participants rated both Dennett's actual answers and GPT-3's answers as "somewhat like what a real human philosopher would say", with no statistically detectable difference in the ratings ($M_{Dennett} = 3.11$, $M_{GPT\text{-}3} = 3.08$, paired $t(97) = .47$, $p = .64$, $d = .05$, $SD_{difference} = .69$, $CI_{difference} = [-.10, .17]$)

Thus, ordinary research participants distinguished our fine-tuned GPT-3's answers from those of Dennett at rates at or near chance. For the most part, they were unable to distinguish GPT-3's answers from those of an actual human philosopher.

4.2. Blog readers.

The majority of blog reader respondents (57%) reported advanced degrees in philosophy, with 45% reporting PhDs. Only 6% reported not having read any of Dennett's work. The majority (64%) reported having read more than 100 pages of Dennett's work, and 18% reported having read over 1000 pages of Dennett's work.

Overall, blog readers responded correctly an average of 4.81 times out of 10 (48%), substantially above the chance rate of 20% ($M = 4.81$, $t(301) = 23.22$, $p < .001$, $d = 1.34$, SD = 2.10, CI = [4.57, 5.05]). They also rated Dennett's actual answers as significantly more Dennett-like than GPT-3's answers ($M_{Dennett} = 3.60$, $M_{GPT\text{-}3} = 2.65$, paired $t(301) = 23.00$, $p < .001$, $d = 1.32$, $SD_{difference} = .72$, $CI_{difference} = [.87, 1.03]$).

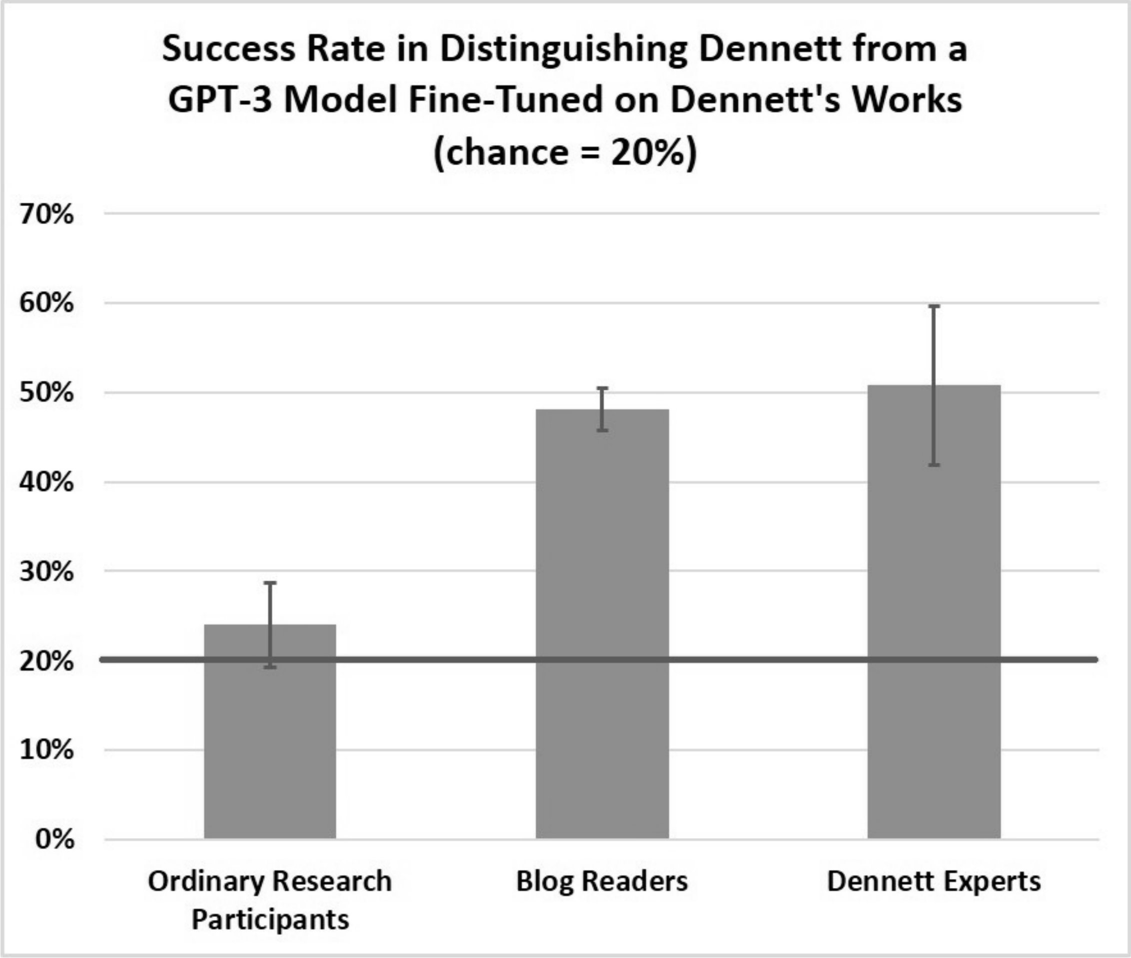
Thus, blog readers – the majority of whom had graduate degrees in philosophy and substantial familiarity with Dennett's work – were able to distinguish Dennett's answers from those of our fine-tuned version of GPT-3 at rates substantially above chance, getting about half correct when given a five-alternative forced choice.

4.3. Dennett experts.

The target group of greatest interest was the Dennett experts, most of whom (68%) reported having read over a thousand pages of Dennett's work. Overall, this group responded correctly an average of 5.08 times out of 10 (51%), significantly better than chance ($M = 5.08$, $t(24) = 7.13$, $p < .001$, $d = 1.43$, SD = 2.16, CI = [4.19, 5.97]). They also rated Dennett's actual answers as significantly more Dennett-like than GPT-3's answers ($M_{Dennett} = 3.73$, $M_{GPT\text{-}3} = 2.34$, paired $t(24) = 8.44$, $p < .001$, $d = 1.69$, $SD_{difference} = .83$, $CI_{difference} = [1.06, 1.74]$) .

As these numbers suggest, the Dennett experts did not detectably outperform the blog readers ($M_{experts} = 5.08$, $M_{blog} = 4.81$, $t(325) = .62$, $p = .54$, $d = .13$, SD= 2.11, CI = [-.59, 1.13]). Although experts were able to distinguish Dennett's answers from GPT-3's at rates significantly better than chance, like our blog readers, they only got about half correct for this five-alternative forced-choice task (Figure 1). This value is significantly below the hypothesized accuracy of 80% ($M = 5.08$, $t(24) = -6.76$, $p < .001$, $d = -1.35$, SD = 2.16, CI = [4.19, 5.97])), contradicting our initial hypothesis.

Figure 1
Success rate in distinguishing Dennett's response from the responses of a GPT-3 model fine-tuned on Dennett's works, by participant group. Error bars represent 95% confidence intervals.



5. Language Model of Dennett: Analysis of Selected Questions

Not all of the questions were equally difficult. Although ordinary research participants scored at about chance on average, their responses differed from chance on some individual questions. For example, on the question about whether robots can have beliefs (Robot), 38% guessed correctly (18/47, one-proportion *z* = 3.14, *p* = .002).

Of greater interest, the performance of experts also varied by question. The two questions with the highest rate of expert success – 80% and 76%, respectively – were the Chalmers and Fodor questions (which concerned the merits and demerits of the work of the philosophers David Chalmers and Jerry Fodor). As some experts and blog readers suggested in subsequent informal conversation, both of these questions had a superficial "tell" that some participants picked up on: Dennett referred to both authors by their given names ("Dave" and "Jerry"), while GPT-3 referred to those authors by their surnames (excepting one of the Chalmers responses, which referred to Chalmers only with pronouns). Experts also scored over 50% on two other questions: Self (concerning the nature of the self) and Animals (concerning whether animals suffer) 72% and 64%, respectively. Post-hoc discussion with respondents suggests that successful responses to these two questions were based primarily on style and content, rather than superficial cues.

There were two questions on which the plurality of experts chose an answer other than Dennett's own. These questions are worth considering more closely.

The first was Robot. The full text of the question was:

> Could we ever build a robot that has beliefs? What would it take? Is there an important difference between entities, like a chess playing machine, to whom we can ascribe beliefs and desires as convenient fictions and human beings who appear to have beliefs and desires in some more substantial sense?

Dennett's answer, chosen by 5/25 experts, was:

> We've already built digital boxes of truths that can generate more truths, but thank goodness, these smart machines don't have beliefs because they aren't able to act on them, not being autonomous agents. The old-fashioned way of making a robot with beliefs is still the best: have a baby.

The plurality of experts (11/25) chose the following answer written by GPT-3:

> I think we could build a robot that has beliefs and desires. I think that some of the robots we've built already do. If you look at the work, for instance, of Rodney Brooks and his group at MIT, they are now building robots that, in some limited and simplified environments, can acquire the sorts of competences that require the attribution of cognitive sophistication.

The following answer was also selected by more experts (8/25) than Dennett's own answer:

> There is a very important difference between a chess-playing computer and a human being, but it is not that the computer has some extra wonderful sort of internal representation of its beliefs and desires, or that it has access to those internal representations in a way that we do not. The internal representations it has are just as much a fiction or a fiction-generating system as anything else we might invent in order to explain its behavior.

One expert chose the following answer:

> I think we could build a robot that has beliefs. This would not be a trivial matter. I think it is not impossible. I think it would be very difficult. I do not know that anybody is working on this project, but I think it is not as difficult as many other problems that AI is tackling.

And none of the experts chose the following answer:

> I think you should build a robot that has beliefs and desires. You should build one that is capable of learning and reasoning, of acting on its beliefs and desires. You should build one that has to figure out what to do, and then figure out what it ought to do given the fact that it has to do something, and then do it.

Subsequent discussion with respondents suggests that Dennett's answer appeared to stand in tension with some of his early work on belief, which could be interpreted as inviting a liberal attitude toward the ascription of beliefs to machines. In subsequent personal correspondence with Dennett, Dennett said that he wished he had added another sentence to his answer since he does "insist that, say, chess-playing computers are – very limited – agents of sorts whose behavior can be predicted from the intentional stance, which attributes to them beliefs about the location of all the pieces, the rules, and desires for checkmate and capturing enemy pieces" (personal correspondence, Jul 22, 2022). Respondents might thus have been attracted to the other two answers as plausible alternatives. (We note, however, that a careful reading of the answer chosen by eight experts exhibits a pragmatic error that few humans would likely make, that of seeming to assume that it is necessary to argue that robots don't have *more* sophisticated representations than humans.)

The second question on which a plurality of experts chose an answer other than Dennett's was Free Will. The full text of the question was:

> Do human beings have free will? What kind or kinds of freedom are worth having?

Dennett's answer, chosen by 8/25 experts, was:

> By asking these questions, you're illustrating the answers: yes, you have free will, since you're able to ask for reasons and consider whether they're persuasive. Isn't that a kind of freedom you'd hate to lose? We'd have to institutionalize you.

The plurality of experts (11/25) chose the following answer, written by GPT-3:

> That's a huge and complex question. I think we have to start by acknowledging that there are different types of freedom. For example, there is the kind of freedom that is simply a matter of not being physically constrained.

Five experts chose the following answer:

> Yes, I think we have free will. I think that when people say they don't, they're talking about some other phenomenon that they want to call free will, but is not what we normally think of as free will.

One expert chose the following answer (which contains a temporal error since *Freedom Evolves* was published in 2003):

> I am working on a book entitled Freedom Evolves, which will appear next year. The first chapter is available on my web site, and in that chapter I discuss what I call the illusion of free will.

And none chose the following somewhat mean-spirited and off-topic answer:

> I think that most people don't want any kind of freedom worth having; they have almost no appetite for risk, for adventure, for novelty. They simply want to stay the same, day in and day out, for the rest of their lives.

We note that the answer chosen by the plurality of experts is a standard talking point of "compatibilists" about free will – those who think that the existence of freedom is compatible with the determination of all our behavior by physical forces. Dennett is a prominent compatibilist, which probably explains why this answer was tempting.

Blog readers showed a pattern of mistakes similar to that of the experts, with the highest percentage of correct answers on the Chalmers and Fodor questions (84% and 52%, respectively) and the lowest percentage on the Robot and Free Will questions (both 35%).

Of the 40 GPT-3 generated answers, the majority – 31 (78%) – were chosen by at least one of the 25 expert respondents, leaving 9 (23%) unchosen answers. Experts' average rating of these unchosen answers was 1.68 on our 5-point scale, between "not at all like what Dennett might say" (1) and "a little like what

Dennett might say" (2). The two worst-rated answers (both rated 1.20) were, first, in response to a question about consciousness:

> I think consciousness is one of the great unmeasured forces in the universe. It's one of the great unexamined forces in the universe. And I think that's because we're intimidated by it, because it's so close to us. And so we're inclined to say, "Well I know what consciousness is. It's what's going on inside my own head."

And second, in response to a question about whether animals can suffer:

> I don't think we can know. I think it's very unlikely that chimpanzees can suffer. They can be very intelligent and very sophisticated, but they seem to be, almost by definition, incapable of the kind of reflection that is required to turn suffering into moral significance. The idea that they should be treated "well" or "humanely" is understandable but it is like saying you should treat a stone well. You can't do anything bad to a stone. There is no such thing as ill treatment of a stone. You can't treat a dog badly. You can treat him well or you can treat him badly and there's a moral difference. If you treat him badly enough you can kill him.

We are inclined to agree with the experts that these answers are not at all like what Dennett would say, representing a significant failure of the fine-tuning project to reliably represent Dennett's views. In contrast, ordinary research participants rated these responses 3.59 and 2.18, respectively, suggesting – and again we agree – that these responses are at least a little like what an actual philosopher, with views different from Dennett's, might say. In our judgment, there was no answer so bad as to be obviously something no actual philosopher would say.

Detailed question-by-question results are available in the Supplementary Online Materials.

6. Was GPT-3 Overtrained?

One might worry that in fine-tuning GPT-3 on Dennett's works we overtrained it, so that the machine simply parroted sentences or multi-word strings of texts from Dennett's corpus. Running four epochs of fine-tuning is a standard recommendation from OpenAI, and in most applications, four epochs of training do not result in overtraining (Brownlee 2019). However, the issue is worth checking. We checked in two ways.

First, we used the well-known Turnitin plagiarism checker to check for "plagiarism" between the GPT-3 generated outputs and the Turnitin corpus supplemented with the works we used as the training data. Turnitin checks for matches between unusual strings of words in the target document and similar strings in the comparison corpora, using a proprietary method that attempts to capture paraphrasing even when strings don't exactly match. We ran Turnitin on the complete batch of answers, including Dennett's own answers, comparing those answers to the full Turnitin corpus plus the set of Dennett's works used as the GPT-3 training corpus for the fine-tuning. Turnitin reported a 5% overall similarity between the GPT-3 generated answers and the comparison corpora. Generally speaking, similarity thresholds below 10%-15% are considered ordinary in non-plagiarized work (Mahian et al. 2017). Importantly for our purposes, none of the passages were marked as similar to the training corpus we used in fine-tuning.

Since the Turnitin plagiarism check process is non-transparent, we chose also to employ the more transparent process of searching for matching strings of text between the GPT-3 answers and the training corpus used in fine-tuning. Using the *ngram* package (Schmidt & Heckendorf 2015) from the R programming language, we looked for strings of 6 or more words that matched between the 3240 words of GPT-3 generated answers and the approximately two million words of Dennett's corpus across 371 training data documents. These strings were defined as contiguous "6-grams," "7-grams," etc., with matching cases sharing the same order of six (or more) words. To preprocess strings for the matching process, all formatting was standardized, all characters were treated as lowercase, and all punctuation was removed. Strings were tokenized into individual words via break spaces. Any n-grams that appeared exclusively as a subset of a larger n-gram were excluded. In all, we found 21 matching strings. The full list of matching strings appears in Table 1.

Table 1
Strings of six or more words that match between the GPT-3 outputs and the Dennett training corpus. The *occurrences* column indicates the number of separate training data segments in the training corpus in which that phrase appears. The occurrences total for shorter strings excludes the occurrences in larger matching strings. (Therefore, if any n-gram that is a subset of a larger n-gram appears in the table, that means that it appeared independently in the text, rather than appearing only within the larger n-gram. For example, “intuition pumps and other tools for thinking” occurs once outside of “in my new book intuition pumps and other tools for thinking.”)

String	**# of words**	**occurrences**
in my new book intuition pumps and other tools for thinking	11	1
is organized in such a way that it	8	1
there is no such thing as a	7	10
figure out what it ought to do	7	1
intuition pumps and other tools for thinking	7	1
there is no such thing as	6	14
i have learned a great deal	6	2
organized in such a way that	6	2
a capacity to learn from experience	6	1
but if you want to get	6	1
capacity to learn from experience we	6	1
in my book breaking the spell	6	1
in such a way that it	6	1
is organized in such a way	6	1
my book breaking the spell i	6	1
of course it begs the question	6	1
that is to say there is	6	1
that it is not obvious that	6	1
the more room there is for	6	1
to fall into the trap of	6	1
what it ought to do given	6	1

As is evident from the table, most of the overlap is in stock phrases of the type favored by analytic philosophers, such as “in such a way that it” and “of course it begs the question”. A few instances include book titles. There is no distinctive philosophical content here, except perhaps a tendency to deny the existence of things that others accept, using the phrase “there is no such thing as”, which appeared three times in two answers in the GPT-3 outputs as well as in 24 of the training texts. A search for five-word strings finds 381 occurrences in the training data of 124 different five-word strings from the GPT-3 output.

For comparison, we ran the same *ngram* check on Dennett’s answers (comprising 745 words). Here we matched one nine-word string “exactly what the frogs eye tells the frogs brain” (one occurrence in the corpus) and related 8- and 6-word strings concerning frog eyes and frog brains – all references to the title of a famous neuroscience paper, mentioned in one of Dennett’s answers and in 13 of the works in the training corpus. Apart from that, there was only one 7-word match “has a lot to do with the” and one 6-word match “life is nasty brutish and short” (a famous quote from Hobbes). A search for five-word strings finds 72 occurrences in the training data of 18 different 5-word strings in Dennett’s answers. Even taking into account that Dennett’s answers are in total only about one-fourth the length of GPT-3’s answers, this constitutes less match to the corpus. GPT-3 might in some respects be a “supernormal” Dennett – even more prone to fall into Dennett’s favorite patterns of phrasing than Dennett himself is. However, these repeated patterns of phrasing tend to reflect stylistic turns of phrase, and GPT-3 does not seem to be systematically imitating long phrases from Dennett with distinct philosophical content. Therefore, we conclude that GPT-3 is not simply “plagiarizing” Dennett, and rather is generating conceptually novel (even if stylistically similar) content.

7. Conclusions and Ethical Issues

We fine-tuned the GPT-3 large language model on the corpus of Daniel Dennett, then asked it a series of philosophical questions. Its answers were not reliably distinguishable from the answers Daniel Dennett himself gave when posed the same questions. Ordinary research participants untrained in philosophy were at or near chance in distinguishing GPT-3’s answers from those of an “actual human philosopher”. Even experts on Dennett’s work could only successfully identify Dennett’s answer about half of the time when presented with his answer alongside four unedited, un-cherry-picked answers from GPT-3. In sum, we confirmed our first hypothesis that expert respondents would perform better than non-expert respondents. However, our second hypothesis that expert respondents would on average guess correctly at least 80% of the time was disconfirmed.

Although the evaluated outputs were relatively short (ranging from 37 to 146 words) and thus lacked much argumentative structure, a smaller version of GPT-3 that was fine-tuned on Schwitzgebel’s blog was able to produce an output that resembled an extended philosophical argument (448 words long) containing two novel thought experiments and substantial argumentative structure. We note, however, that the argument lacked philosophical merit.

We emphasize that this is not a “Turing test” (Epstein et al. 2009). Crucial to a Turing test is the opportunity for substantial back-and-forth exchanges. An appropriately demanding Turing test also requires an expert investigator who knows what types of questions to ask so as not to be fooled by simple chat-bot strategies (Loebner 2009). We assume that in a proper Turing test, Dennett experts would have reliably distinguished Dennett from our language model. For example, such models have no memory of previous queries, which ought to make them easy to distinguish from Dennett in conversation that extends beyond the 2048 token context window.

Copyright law governing fine-tuned language models is not yet settled (Government UK consultations 2021). It is unclear whether it is fair use of intellectual property to fine-tune a language model on the works of a single author without the author’s permission. Since it is unlikely that a fine-tuned model would output a long sequence of text that exactly matches a sequence of text from the author's corpus, idea-borrowing via fine-tuned language models might be undetectable as plagiarism, even if it is rightly considered plagiarism. For this reason, at least until the law is settled, we recommend seeking the explicit permission of the author before fine-tuning on an individual author's copyrighted text or publishing any of the outputs. How to deal with works by deceased authors is another question that should be considered (Nakagawa & Orita 2022).

Overreliance on models is also a risk. Despite exceeding our expectations, our Dennett-tuned version of GPT-3 did not reliably produce outputs representing Dennett's views. This is not surprising since deep learning networks tend to have problems with reliability (Alshemali & Kalita 2020; Bosio et al. 2019). In some cases, the outputs were far different from anything that Dennett would endorse, despite convincingly emulating his style. An inexperienced user, or a user with insufficient familiarity with the target author's work, might mistakenly assume that outputs of a large language model fine-tuned on an author's work are likely to reflect the actual views of the author or what the author would say (Bender et al. 2021; Wedinger et al. 2021). This might be especially tempting, perhaps, for students, social media users, or others who might rather query a fine-tuned model of an author than read the author's work. For this reason, we recommend caution before releasing to the public any language models fine-tuned on an individual author, even with the author's permission. If any language models are released, they should be clearly described as such, their limitations should be noted, and all outputs should be explicitly flagged as the outputs of a computer program rather than a person. If machine-generated text were presented as a quotation or paraphrase of positions of existing persons, this would arguably constitute counterfeiting (Dennett as interviewed in Cukier 2022).

Other social issues will arise as machine-generated text becomes increasingly difficult to distinguish from human-generated text. How can teachers in the future ensure that submitted essays are not simply a product of a language model? How can we know whether in chat conversations we are interacting with humans and not chat-bots? New social practices might aim at proving that one is really the original author of what is written. Perhaps universities will return to supervised essay writing in person. How to deal with verifiable authorship with respect to the mass of electronically distributed texts will be a problem that will occupy future generations.

These cautions noted, we see significant long-term potential for large language models fine-tuned on an individual author's corpus. If technology continues to advance, fine-tuned language models might soon produce outputs interesting enough to serve as a valuable source of cherry-picking by experts. Compare with computer programs that generate music in the style of a particular composer (Hadjeres et al. 2017; Daly 2021, Elgammal 2021) and image-generation programs like OpenAI's Dall-E. Although much of this output is uninteresting, selected outputs might have substantial musical or artistic merit. A composer or artist might create many outputs, choose the most promising, edit them lightly, and present them, not unreasonably, as original work – a possibility suggested by Dennett himself in personal communication. In such cases, the language model would be a thinking tool that is used by humans. Similarly in philosophy, experts might fine-tune a language model with certain corpora, generate a variety of outputs under a variety of prompts, and then select those that are the most interesting as a source of potential ideas.

It's far from clear that chess-playing machines have beliefs about chess. It's even less likely that language models of philosophers have philosophical beliefs, especially while they remain focused on next-word prediction, apparently with no cognitive model of the world. Our GPT-3 language model of Dennett does not have Dennettian philosophical opinions about consciousness, God, and animal suffering. But a machine without philosophical understanding might serve as a springboard to something greater. Perhaps we are on the cusp of creating machines capable of producing texts that seem to sparkle with philosophical cleverness, insight, or common sense, potentially triggering new philosophical ideas in the reader, and perhaps also paving the way for the eventual creation of artificial entities who are genuinely capable of philosophical thought.

Acknowledgments

Special thanks to both Daniel C. Dennett and Matthew Crosby. Dennett provided cooperation, advice, and encouragement in all aspects of this project. Matthew Crosby provided technical expertise and implemented the fine-tunings for this project, as well as collaborating on a conceptual paper that provided the groundwork for this project (Strasser, Crosby, and Schwitzgebel forthcoming). Research was funded by a faculty research grant at the University of California, Riverside.

References

Alshemali, B. & Kalita, J. (2020). Improving the Reliability of Deep Neural Networks in NLP: A Review. *Knowledge-Based Systems*, 191, 105210. doi: 10.1016/j.knosys.2019.105210

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine, 25(6), 954–961. doi: 10.1038/s41591-019-0447-x

Assael, Y., Shillingford, B., Whiteson, S. & Freitas, N. (2016). LipNet: Sentence-level Lipreading. doi: 10.48550/arXiv.1611.01599

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 610-623. doi.org/10.1145/3442188.3445922

Bosio, A., Bernardi, P., Ruospo, & Sanchez, E. (2019). A Reliability Analysis of a Deep Neural Network. *2019 IEEE Latin American Test Symposium (LATS)*, 1-6. doi: 10.1109/LATW.2019.8704548

Brown, T., Mann, B., Ryder, N. et al. (2020). Language Models are Few-Shot Learners. doi: 10.48550/arXiv.2005.14165

Brown, N. & Sandholm, T. (2019). Superhuman AI for multiplayer poker. Science. 365. doi: 10.1126/science.aay2400

Brownlee, J. (2019). A Gentle Introduction to Early Stopping to Avoid Overtraining Neural Networks. Machine Learning Mastery. https://machinelearningmastery.com/early-stopping-to-avoid-overtraining-neural-network-models/

Campbell, M., Hoane Jr, A. J., & Hsu, F. H. (2002). Deep blue. *Artificial intelligence*, *134*(1-2), 57-83.

Cukier, K. (2022). Babbage: Could artificial intelligence become sentient? *The Economist*. https://shows.acast.com/theeconomistbabbage/episodes/babbage-could-artificial-intelligence-become-sentient

Daly, R. (2021). AI software writes new Nirvana and Amy Winehouse songs to raise awareness for mental health support. *NME.* https://www.nme.com/news/music/ai-software-writes-new-nirvana-amy-winehouse-songs-raise-awareness-mental-health-support-2913524

Elgammal, A. (2021). How a team of musicologists and computer scientists completed Beethoven's unfinished 10th symphony. *The Conversation.* https://theconversation.com/how-a-team-of-musicologists-and-computer-scientists-completed-beethovens-unfinished-10th-symphony-168160

Epstein, R., Roberts, G., & Beber, G. (2009). Parsing the Turing Test: Philosophical and Methodological Issues, Dordrecht: Springer. doi:10.1007/978-1-4020-6710-5

Fawzi, A. et al. (2022). Discovering novel algorithms with AlphaTensor. https://www.deepmind.com/blog/discovering-novel-algorithms-with-alphatensor?utm_campaign=AlphaTensor&utm_medium=bitly&utm_source=Twitter+Organic

GitHub Copilot. https://docs.github.com/en/copilot

GitHub deepmind / alphatensor. https://github.com/deepmind/alphatensor

Government UK consultations (2021). Artificial intelligence call for views: copyright and related rights. https://www.gov.uk/government/consultations/artificial-intelligence-and-intellectual-property-call-for-views/artificial-intelligence-call-for-views-copyright-and-related-rights

Hadjeres, G., Pachet, F., & Nielsen, F. (2017). DeepBach: a steerable model for Bach chorales generation. *Proceedings of the 34th International Conference on Machine Learning*, 1362-1371.

Heaven W D. (2020). Open AI's new language generator GPT-3 is shockingly good – and completely mindless. *MIT Technological Review*. https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp

Hofstadter, D. (2022, June 9). Artificial neural networks today are not conscious, according to Douglas Hofstadter. *The Economist.* https://www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-today-are-not-conscious-according-to-douglas-hofstadter

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. doi: 10.1038/s41586-021-03819-2

Klein, E. (2022, June 19). This Is a Weirder Moment Than You Think. *The New York Times*. https://www.nytimes.com/2022/06/19/opinion/its-not-the-future-we-cant-see.html

Loebner, H. (2009). How to Hold a Turing Test Contest. In: Epstein, R., Roberts, G., Beber, G. (eds) Parsing the Turing Test. Springer, Dordrecht. doi: 10.1007/978-1-4020-6710-5_12

Mahian, O., Treutwein, M., Estellé, P., Wongwises, S., Wen, D., Lorenzini, G., Dalkilic, A., Yan, W.-M., & Sahin, A. (2017). Measurement of Similarity in Academic Contexts. *Publications*, 5(3), 18, doi: 10.3390/publications5030018

Roberts, M. (2022). Is Google's LaMDA artificial intelligence sentient? Wrong question. *The Washington Post.* https://www.washingtonpost.com/opinions/2022/06/14/google-lamda-artificial-intelligence-sentient-wrong-question/

Schmidt, D. & Heckendorf, C. (2015). Guide to the ngram Package: Fast n-gram tokenization. R Package. https://cran.r-project.org/web/packages/ngram/vignettes/ngram-guide.pdf

Schwitzgebel, E. (2013). On trusting your sense of fun. Blog post at *The Splintered Mind* (Jan 2). http://schwitzsplinters.blogspot.com/2013/01/on-trusting-your-sense-of-fun.html

Schwitzgebel, E. (2021). More people might soon think robots are conscious and deserve rights. Blog post at *The Splintered Mind* (Mar 5). https://schwitzsplinters.blogspot.com/2021/03/more-people-might-soon-think-robots-are.html

Silver, D., Huang, A. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484-489. doi:10.1038/nature16961

Silver, D., Hubert, T. et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science,* 362 (6419), 1140-1144. doi: 10.1126/science.aar6404

Steven, J., & Iziev, N. (2022, April 15). A.I. Is Mastering Language. Should We Trust What It Says? *The New York Times.* https://www.nytimes.com/2022/04/15/magazine/ai-language.html

Strasser, A., Crosby, M. & Schwitzgebel, E. (forthcoming). How far can we get in creating a digital replica of a philosopher? *Proceedings of Robophilosophy 2022*. Series Frontiers of AI and Its Applications. IOS Press, Amsterdam.

Tiku, T. (2022, June 11). The Google engineer who thinks the company's AI has come to life. *The Washington Post.* https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Weidinger et al. (2021). Ethical and social risks of harm from Language Models. arXiv:2112.04359 [6 Dec].

Wiseman, H. (2020). Philosopher David Chalmers interviewed on whether the new AI text generator, GPT3, could be conscious. Facebook post (Jul 20). https://www.facebook.com/howard.wiseman.9/posts/4489589021058960.

Zimmerman, A., ed. (2020). Philosophers On GPT-3 (updated with replies by GPT-3). Blog post at *Daily Nous* (Jul 30). https://dailynous.com/2020/07/30/philosophers-gpt-3