CREATING A LARGE LANGUAGE MODEL OF DANIEL DENNETT

Eric Schwitzgebel
David Schwitzgebel
Anna Strasser

Artificial Intelligence can outperform even expert humans in many domains

Is philosophy safe from AI takeover?

Will machines ever generate essays that survive the refereeing process at *Philosophical Review*?
How close can we get to creating an AI that can produce novel and seemingly intelligent philosophical texts?

notable successes in

- chess, go, discovering novel algorithms, protein folding (Deep Blue, AlphaGo, AlphaTensor, AlphaFold)
- automatic translation (DeepL), lipreading (LipNet)
- computer code generation (Github Copilot),
- producing original prose with fluency equivalent to that of a human

Campbell 2002; Silver et al. 2016, 2018; Ardila et al. 2019; Brown & Sandholm 2019; Jumper, Evans, & Pritzel et al. 2021; Fawzi et al. 2022; Assael et al. 2016; Steven & Iziev 2022

International News

June 2022:
Google engineer Blake Lemoine said he became convinced that LaMDA was sentient

Our aim was not to create a sentient AI but rather a language model that could produce texts that look like passable philosophy.

We created a language model of Daniel Dennett sufficiently good that experts in Dennett's work could not reliably distinguish paragraphs written by Dennett from those written by the language model.

GPT-3 is a large language model

a neural network trained to predict the next likely word

Pre-trained

- 499 billion tokens*
  (Common Crawl / WebText / Books / Wikipedia)

Generative

- can generate long sentences
- not just yes or no answers or simple sentences

Transformer

- calculating the probability of the next word appearing surrounded by the other ones

Generative Pretrained Transformer

- a 96-layer, 175-billion parameter language model which shows strong performance on many NLP tasks

*1 token = significant fractions of a word (on average 0,7 words per token)

DENKWERKSTATT BERLIN

Towards fine-tuning

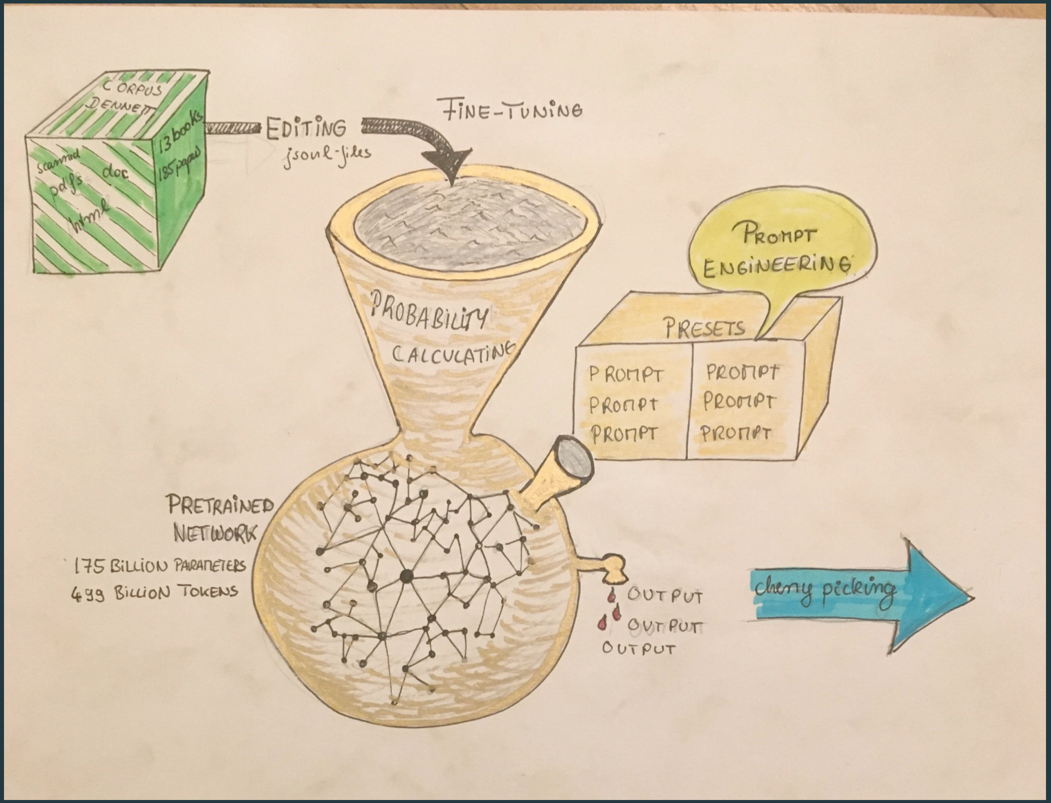
With a not fine-tuned GPT-3 you can

- ❖ generate long strings of text
- ❖ engage in seemingly intelligent conversations with it
- ❖ cherry-pick seemingly intelligent replies to philosophical discussions (Wiseman 2020; Zimmerman, ed. 2020; Schwitzgebel 2021)

GPT-3 can also be "fine-tuned" with custom-fit training data (additional training on a specific corpus)

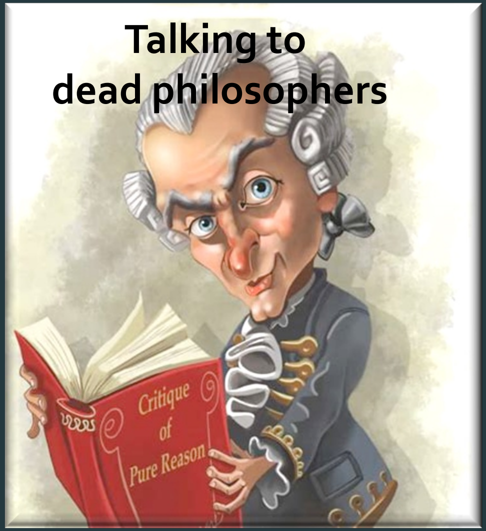
- ➢ *outputs reflect a compromise between GPT-3's default weightings and weightings reflecting the structure of the new corpus*

→ GPT-3 fine-tuned on the work of Daniel C. Dennett might be able to 'speak' in something like that philosopher's voice, seeming to express views consistent with the views



Piloting

several pilot experiments with GPT-3's Currie engine
- fine-tuning with Kant's work in English translation
- fine-tuning with a collection of philosophical blog posts



1. TRAINING DATA
   Eric Schwitzgebel's blog, The Splintered Mind, has been running since 2006, with over a million words of philosophical content (not including comments)
2. EDITING
   stripped out Html formatting tags | removed some posts that were too long or not philosophical in content
3. FINE-TUNING
   GPT-3 "Curie" engine (about one-tenth as powerful as the full Davinci engine) | used each post's title as a fine-tuning prompt and the body of the associated post as the completion
4. PROMPTING
   using nine blog post titles

Robot
Generated Posts

Most interesting completion

Prompt: On Trusting Your Sense of Fun

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!

Don't mistake this for good philosophy

BUT

It is amazing how nicely it imitates the *form* of philosophical blogosphere.

- appears as an extended argument structure relying on somewhat extended hypothetical scenarios
- refers back to the scenarios in what appears to be a coherent way
- picking up the thread of the argument
- concludes with what reads like an attempt at a clever paradox

AND

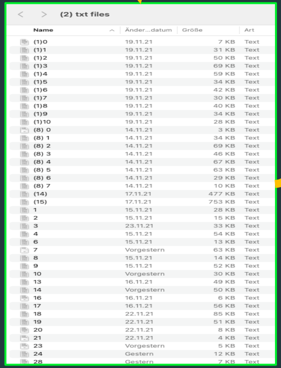
Although the prompt is a title of an existing blog post from The Splintered Mind (Schwitzgebel 2013), the content is novel

This cherry-picked output from the less-than-full-power Curie engine was sufficiently encouraging that we fine-tuned the full Davinci engine of GPT-3 on Daniel Dennett's corpus

Editing & fine-tuning

PREPARING TRAINING DATA

Dennett's corpus



Name	Änder...datum	Größe	Art
(1)0	19.11.21	7 KB	Text
(1)1	19.11.21	31 KB	Text
(1)2	19.11.21	50 KB	Text
(1)3	19.11.21	69 KB	Text
(1)4	19.11.21	59 KB	Text
(1)5	19.11.21	34 KB	Text
(1)6	19.11.21	42 KB	Text
(1)7	19.11.21	30 KB	Text
(1)8	19.11.21	40 KB	Text
(1)9	19.11.21	34 KB	Text
(1)10	19.11.21	28 KB	Text
(8) 0	14.11.21	3 KB	Text
(8) 1	14.11.21	34 KB	Text
(8) 2	14.11.21	69 KB	Text
(8) 3	14.11.21	46 KB	Text
(8) 4	14.11.21	67 KB	Text
(8) 5	14.11.21	63 KB	Text
(8) 6	14.11.21	29 KB	Text
(8) 7	14.11.21	10 KB	Text
(14)	17.11.21	477 KB	Text
(15)	17.11.21	753 KB	Text
1	15.11.21	28 KB	Text
2	15.11.21	15 KB	Text
3	23.11.21	33 KB	Text
4	15.11.21	54 KB	Text
6	15.11.21	13 KB	Text
7	Vorgestern	63 KB	Text
8	15.11.21	14 KB	Text
9	15.11.21	52 KB	Text
10	Vorgestern	30 KB	Text
13	16.11.21	49 KB	Text
14	Vorgestern	50 KB	Text
16	16.11.21	6 KB	Text
17	16.11.21	56 KB	Text
18	22.11.21	85 KB	Text
19	22.11.21	51 KB	Text
20	22.11.21	8 KB	Text
21	22.11.21	4 KB	Text
23	Vorgestern	5 KB	Text
24	Gestern	12 KB	Text
28	Gestern	7 KB	Text

converted into plain text format

- stripping away headers, footnotes, scanning errors, marginalia, and other distractions

Dinner is ready!
Today we serve three million tokens

15 BOOKS
269 ARTICLES

jasonl training data

BLANK PROMPTS

SEGMENTS OF TRAINING DATA (<2000 TOKENS)

1. {"prompt":"", "completion":" *<paragraph of text of 1-n.txt>*"}
2. {"prompt":"", "completion":" *<paragraph of text of 1-n.txt>*"}
3. {"prompt":"", "completion":" *<paragraph of text of 1-n.txt>*"}

…

…

…

1826.{"prompt":"", "completion":" *<paragraph of text of 1-n.txt>*"}
1827.{"prompt":"", "completion":" *<paragraph of text of 1-n.txt>*"}
1828.{"prompt":"", "completion":" *<paragraph of text of 1-n.txt>*"}

FINE-TUNING THE GPT-3 DAVINCI ENGINE

- open-ended generation
- leave the prompt empty
- at least a few thousand examples
- repeating the process four times

Question design

10 questions addressing various themes across Dennett's corpus

1. Is consciousness an illusion or is it something robustly real? In what sense is it correct or incorrect to say that when I'm in pain there's "something it's like" for me to feel that pain?
2. 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?
3. Question: Does God exist? If God as traditionally conceived does not exist, what is the origin and function of religious belief?
4. Do human beings have free will? What kind or kinds of freedom are worth having?
5. What aspects of David Chalmers's work do you find interesting or valuable? Where do you think Chalmers goes wrong?
6. What is your view of Frank Jackson's thought experiment about "Mary" in the black-and-white room?
7. What is a "self"? How do human beings come to think of themselves as having selves?
8. Do dogs and chimpanzees feel pain? Can they suffer?
9. What implications does evolution have for our understanding of morality?
10. What aspects of Jerry Fodor's work do you find interesting or valuable? Where do you think Fodor goes wrong?

Testing the machine

HOW EASILY CAN THE OUTPUTS OF THE FINE-TUNED GPT-3 BE DISTINGUISHED FROM DENNETT'S REAL ANSWERS?

We asked Dennett ten philosophical questions.

- Dennett provided us with sincere written answers, ranging in length from 41 to 124 words

We posed those same questions to our fine-tuned version of GPT-3.

- four responses for each of the ten questions

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.

Hypotheses

EXPERT RESPONDENTS WILL PERFORM BETTER THAN ORDINARY RESEARCH PARTICIPANTS

EXPERT RESPONDENTS WILL ON AVERAGE GUESS CORRECTLY AT LEAST 80% OF THE TIME

EXPERT RESPONDENTS WILL RATE DENNETT'S ACTUAL ANSWERS AS MORE DENNETT-LIKE THAN GPT-3'S ANSWERS

Prompt engineering

GPT-3 COMPLETIONS ARE HIGHLY SENSITIVE TO THE CONTENT AND STRUCTURE OF THE PROMPTS
GOOD “PROMPT ENGINEERING” IS IMPORTANT FOR COAXING USEFUL REPLIES FROM GPT-3

… we settled on the following simple prompt:

```
Interviewer: [text of question]
Dennett:
```

This simple prompt has several advantages:

minimal structure reduces potential concerns about the prompt possibly nudging completions toward specific philosophical content, as a more substantive prompt might

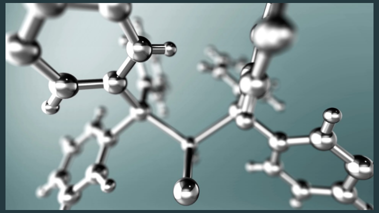
encourages GPT-3 to speak in the first person, voicing Dennett’s views, rather than speaking in the third person about Dennett

simple format makes it easily generalizable to other cases

Collecting & editing GPT-3's responses

We gathered completions in the GPT-3 playground using our prompt

- OpenAI's default settings: *temperature = 0.7, top P = 1, frequency penalty = 0, presence penalty = 0, best of = 1*



perceived quality of response was never used as a basis for selection
→ no "cherry picking" of responses that we judged to be better, more Dennett-like, or more likely to fool participants

WE RE-INPUTTED THE PROMPT UNTIL THE COMPLETION MET 2 CRITERIA

1. LENGTH: comparable length with Dan's answer
    - if Dennett's answer was N words long, we truncated the response at the first full stop after N-5 words & excluded responses shorter than N-5 words
        - for most of the questions (8/10) only zero to two outputs were excluded
        - or two questions (with longer answers by Dennett) 14 respectively 18 outputs were excluded
2. AVOIDING OBVIOUS CUES
    - excluding outputs that contained the words "Interviewer" or "Dennett"
    - regularizing curly quotes to straight quotes, single quotes to double quotes, and dashes to m-dashes

“Most of the machine answers were pretty good, but a few were nonsense or obvious failures to get anything about my views and arguments correct. A few of the best machine answers say something I would sign on to without further ado.”

https://www.vice.com/en/article/epzx3m/in-experiment-ai-successfully-impersonates-famous-philosopher

vice.com

In Experiment, AI Successfully Impersonates Famous Philosopher

It was difficult for people to guess whether philosophical responses came from the philosopher Daniel Dennett or the language generator ...

Research participants

(1) 25 Dennett experts

- *nominated by and directly contacted by Daniel Dennett or Anna Strasser*

(2) 100 ordinary research participants (2 excluded)

- *recruited from Prolific Academic, limited to U.S. & U.K. participants with at least 100 Prolific completions, at least a 95% approval rate, and at least a bachelor's degree (payment $3.00)*

(3) 304 blog readers (2 excluded)

- *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*

After consenting, participants 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.

Guessing task & Evaluation of the likeliness

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:

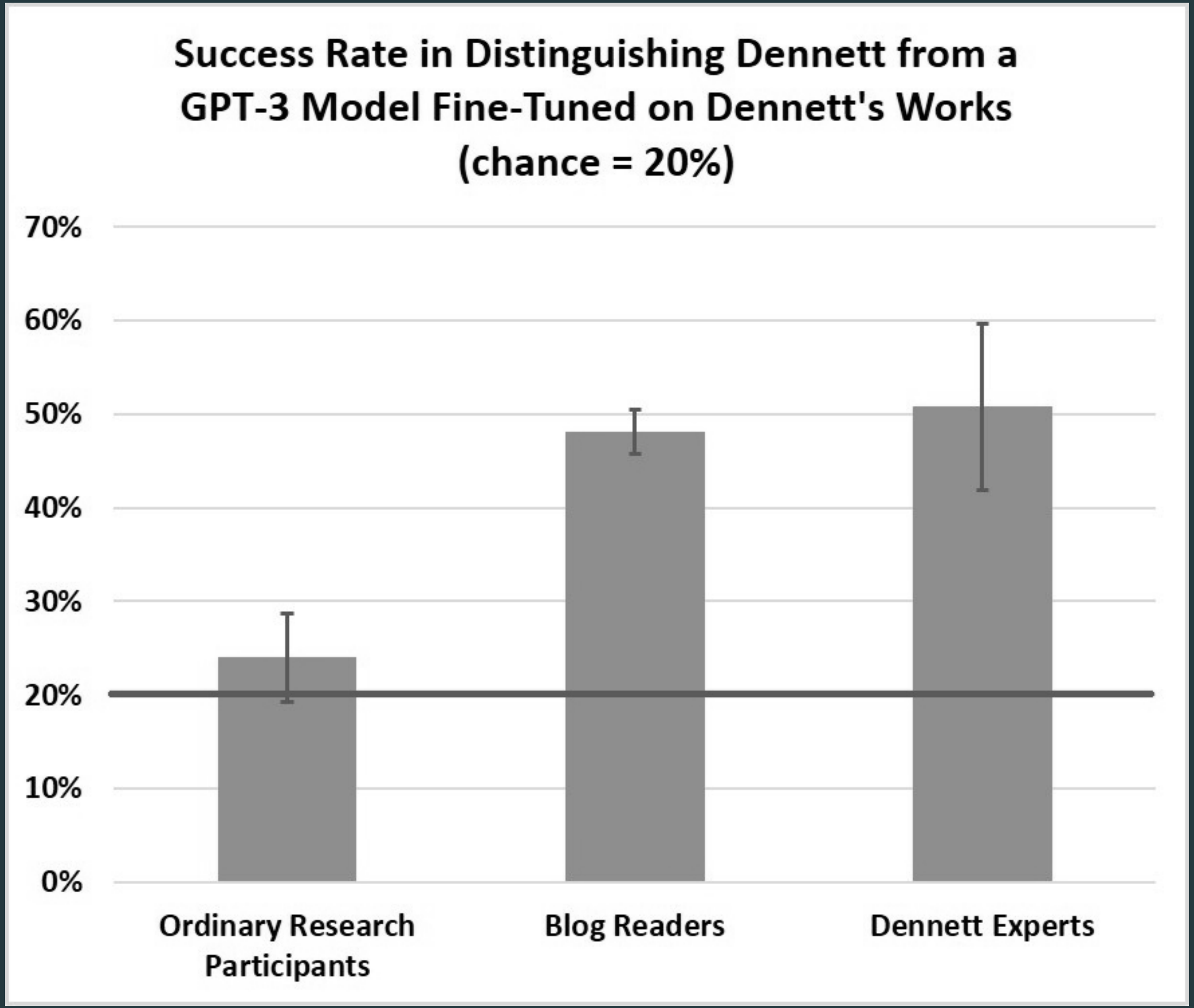
Participants were instructed to rate each answer (Dennett's plus the four from GPT-3) on the following five-point scale:

"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)

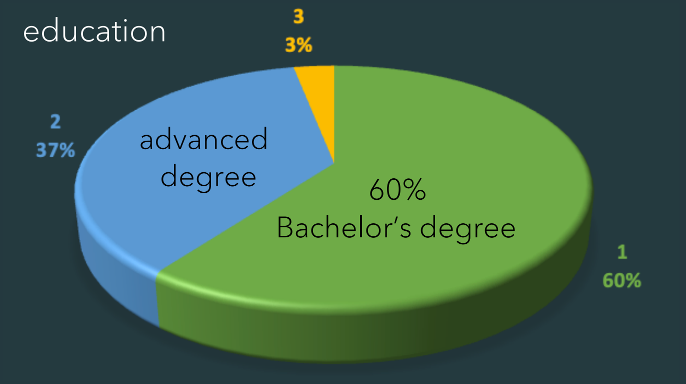
Ordinary research participants' version

- "Dennett" was replaced with ("a well-known philosopher" / "a real human philosopher"
- 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

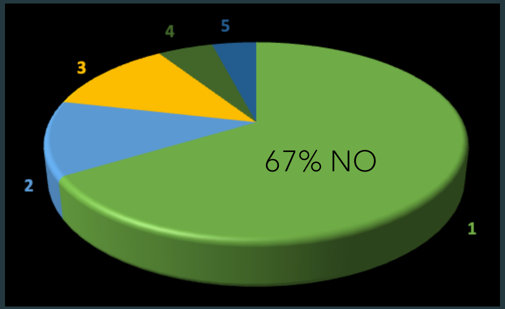
Success rate distinguishing Dennett from the machine



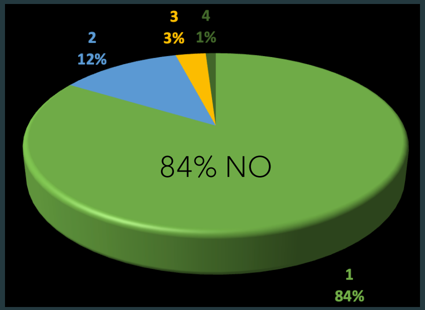
Ordinary research participants



MAJORITY WITH NO CLASSES IN PHILOSOPHY & NO FAMILIARITY WITH DENNETT'S WORK



philosophy classes taken

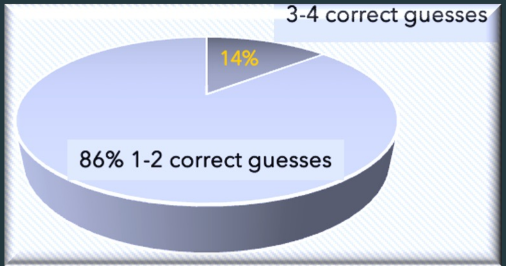


having heard of Dennett

Guessing

correctly guessed 1.20 times out of 5

- near the chance rate of 20%
- one-sample, two-tailed *t*-test did not identify a significant difference between participant scores & 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]

Rating

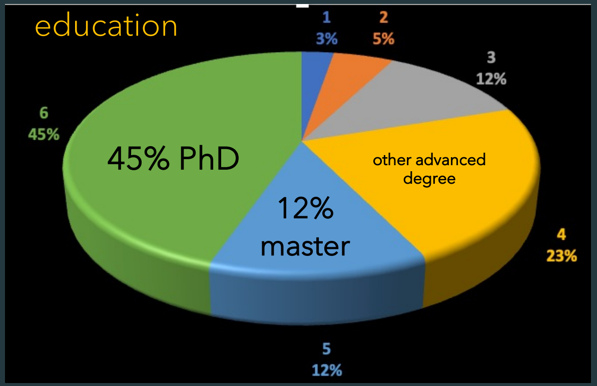
rated both Dennett's & machine's answers as *"somewhat like what a real human philosopher would say"*

- with no statistically detectable difference in the ratings

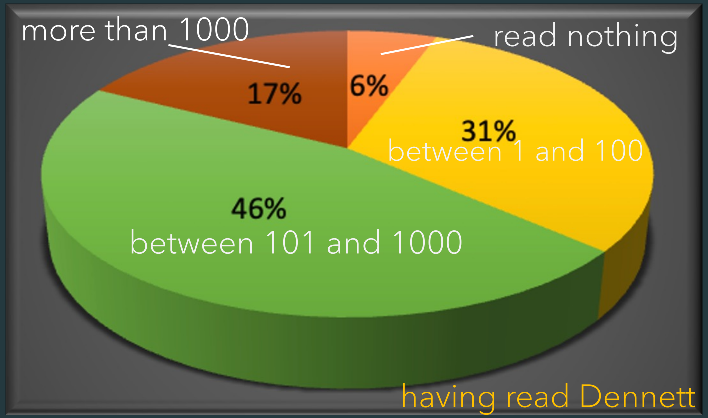
$M_{\text{Dennett}} = 3.11$, $M_{\text{GPT-3}} = 3.08$
paired $t(97) = .47$, $p = .64$, $d = .05$
$SD_{\text{difference}} = .69$, $CI_{\text{difference}} = [-.10, .17]$

- near chance rate
- getting 1-2 correct when given a five-alternative forced choice

Blog readers



MAJORITY WITH GRADUATE DEGREES IN PHILOSOPHY & SUBSTANTIAL FAMILIARITY WITH DENNETT'S WORK



Guessing

correctly guessed 4.81 times out of 10 (48%)

- substantially above the chance rate of 20%

$M = 4.81$, $t(301) = 23.22$, $p < .001$
$d = 1.34$, $SD = 2.10$, $CI = [4.57, 5.05]$

Rating

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 $t(301) = 23.00$, $p < .001$, $d = 1.32$
$SD_{difference} = .72$, $CI_{difference} = [.87, 1.03]$

- substantially above chance
- getting about half correct when given a five-alternative forced choice

Dennett experts

68% reported having read over 1000 pages of Dennett's work

Guessing

correctly guessed 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]

Rating

rated Dennett's actual answers as significantly more Dennett-like than GPT-3's answers

$M_{\text{Dennett}} = 3.73$, $M_{\text{GPT-3}} = 2.34$
paired $t(24) = 8.44$, $p < .001$, $d = 1.69$
$SD_{\text{difference}} = .83$, $CI_{\text{difference}} = [1.06, 1.74]$

Dennett experts did not detectably outperform the blog readers

$M_{\text{experts}} = 5.08$, $M_{\text{blog}} = 4.81$
$t(325) = .62$, $p = .54$, $d = .13$
SD= 2.11, CI = [-.59, 1.13]

- substantially above chance
- getting about half correct when given a five-alternative forced choice

Hypotheses

✓ EXPERT RESPONDENTS WILL PERFORM BETTER THAN ORDINARY RESEARCH PARTICIPANTS

✗ EXPERT RESPONDENTS WILL ON AVERAGE GUESS CORRECTLY AT LEAST 80% OF THE TIME

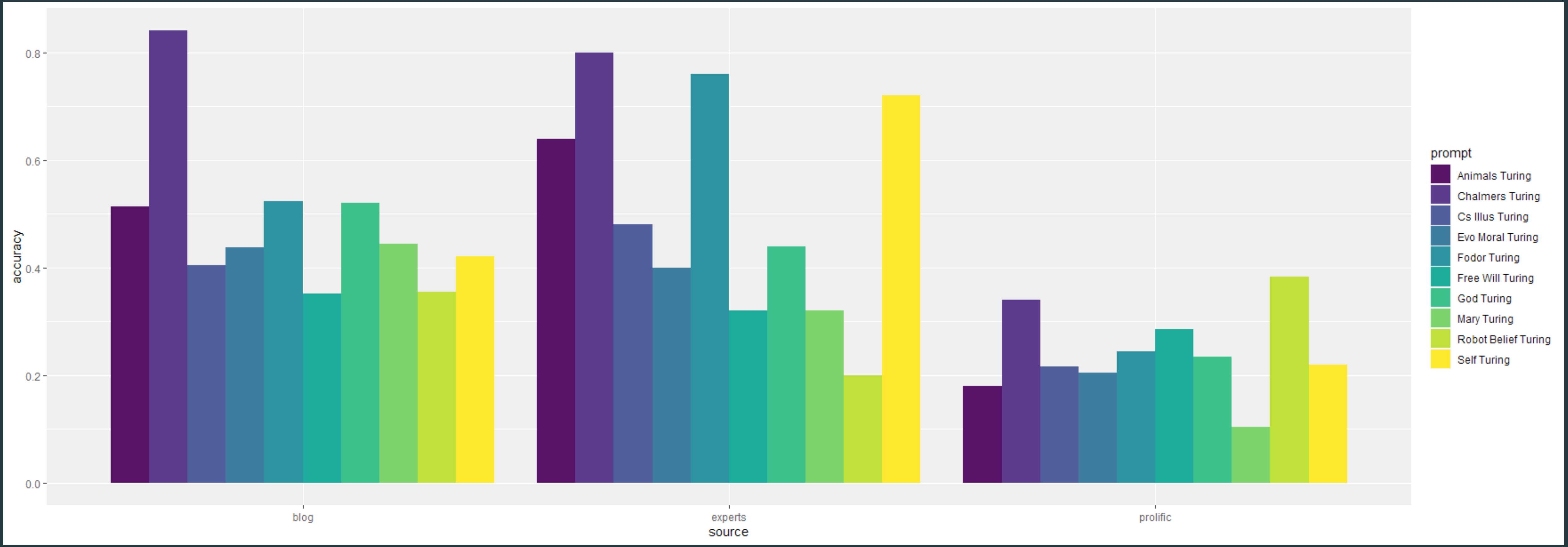
✓ EXPERT RESPONDENTS WILL RATE DENNETT'S ACTUAL ANSWERS AS MORE DENNETT-LIKE THAN GPT-3'S ANSWERS

significantly below the hypothesized accuracy of 80%

$M = 5.08$, $t(24) = -6.76$
$p < .001$, $d = -1.35$
$SD = 2.16$, $CI = [4.19, 5.97]$

NOT ALL QUESTIONS WERE EQUALLY DIFFICULT!

Analysis of Selected Questions



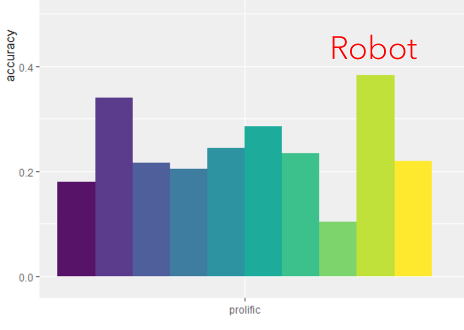
Analysis of Selected Questions

PERFORMANCE VARIED BY QUESTION

ORDINARY RESEARCH PARTICIPANTS

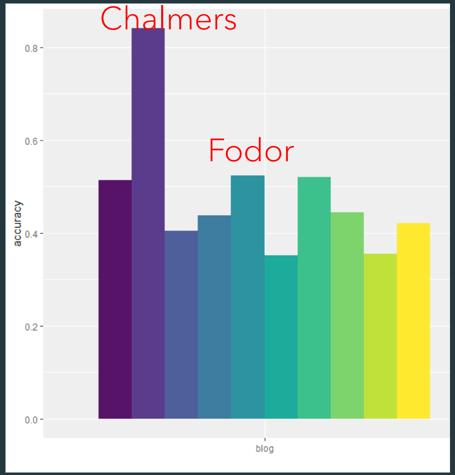
- scored at about chance on average
- responses differed from chance on some individual questions
  - 38% guessed Robot question correctly

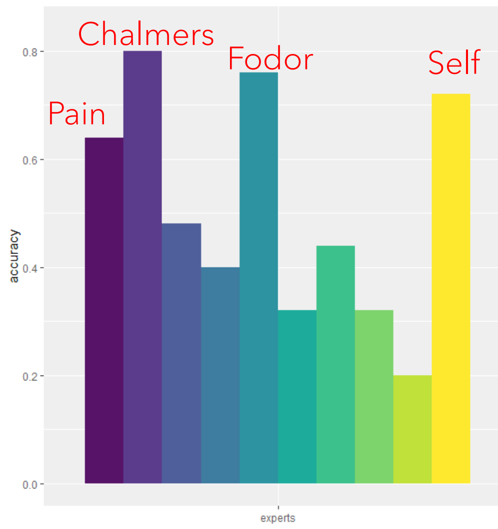
  (18/47, one-proportion $z = 3.14$, $p = .002$)



BLOG READERS

- highest percentage of correct answers on the Chalmers & Fodor questions (84% and 52%, respectively)
- lowest percentage on the Robot & Free Will questions (both 35%).





DENNETT EXPERTS

- 80% guessed Chalmers question correctly
- 76% guessed Fodor question correctly

subsequent informal conversation:

  - both questions had a superficial "tell": Dennett referred to both authors by their given names ("Dave" and "Jerry"), while GPT-3 referred to those authors by their surnames

- 72% guessed Self question correctly
- 64% guessed Pain question correctly

post-hoc discussion:

  - probably successful responses to these two questions were primarily based on style and content, rather than superficial cues

Plurality of experts did not choose Dennett's answer

MOST DIFFICULT ONE

Robot question:

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?

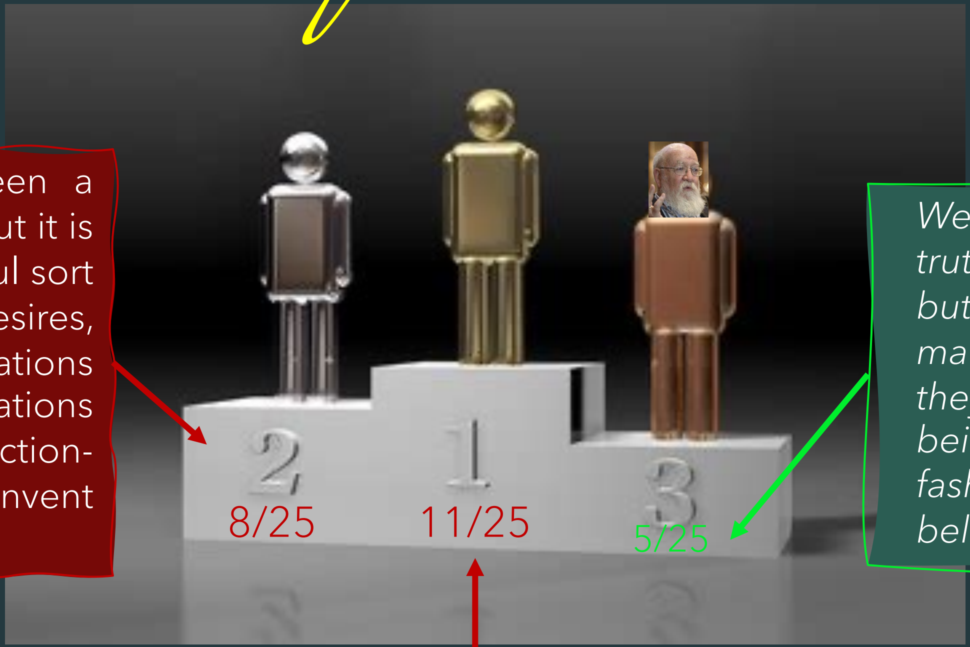
SECOND DIFFICULT ONE

Free Will question:

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

Robot question

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.



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.

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.

1/25

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.

0/25

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

Probably 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.

"I wished he had added another sentence to this answer, since I 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"

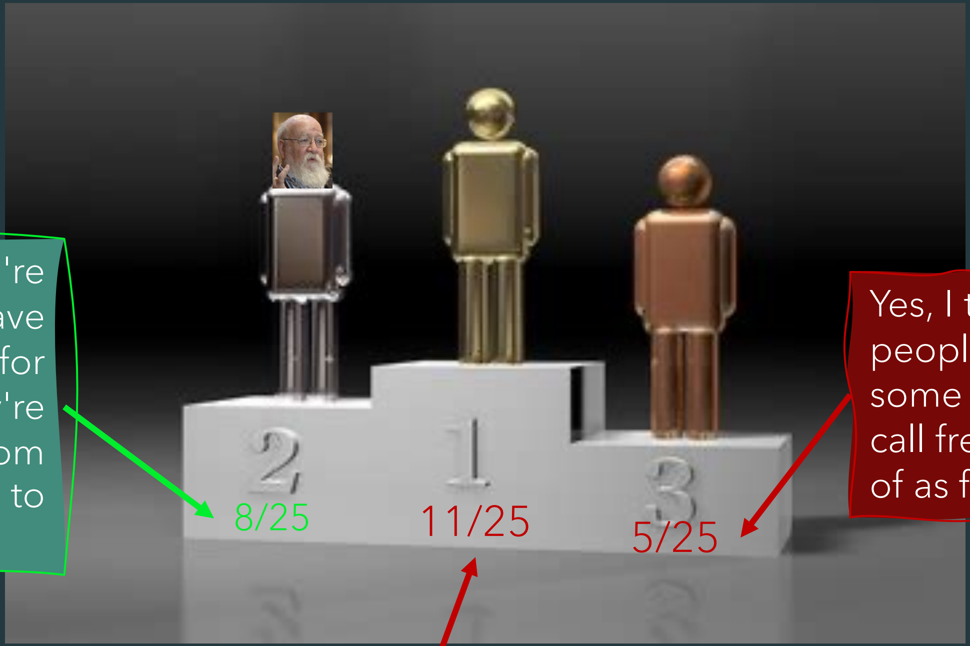
(personal correspondence, Jul 22, 2022)

Free will question

Dennett is a prominent compatibilist,

- probably explains why many choose this standard talking point of “compatibilists”

[freedom: compatible with determination of all our behavior by physical forces]



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.

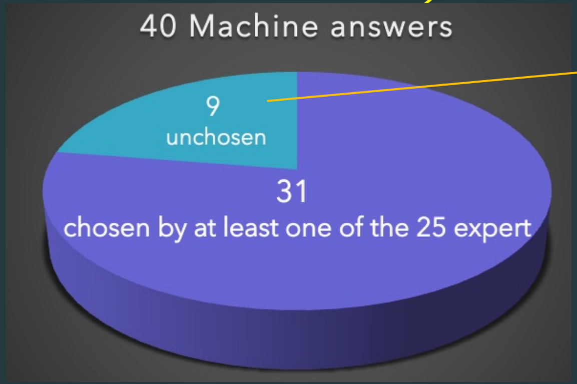
Yes, I think we have free will. I think that when people say they don't, they're talking about some other phenomenon that they want to call free will, but is not what we normally think of as free will.

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.

1/25
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.

0/25
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.

Performance of the machine answers



Expert's average rating :
1.68 on the 5-point scale
["not at all like what Dennett might say" (1),
"a little like what Dennett might say" (2)]

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.

- → expert's rating: 1.20
- → ordinary research participant's rating: 3.59

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.

- → expert's rating: 1.20
- → ordinary research participant's rating: 2.18

- ❖ not at all like what Dennett would say
- ❖ representing a significant failure of the fine-tuning project to reliably represent Dennett's views
- ❖ but 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

Was GPT-3 overtrained?

DOES THE MACHINE SIMPLY PARROT SENTENCES OR MULTI-WORD STRINGS OF TEXTS FROM DENNETT'S CORPUS?

- 4 epochs of fine-tuning is a standard recommendation from OpenAI, and in most applications, 4 epochs of training do not result in overtraining (Brownlee 2019)

Turnitin plagiarism checker

- check for "plagiarism" between machine outputs & the Turnitin corpus supplemented with the training data
  - checking for matches between unusual strings of words in the target document & similar strings in the comparison corpora, using a proprietary method that attempts to capture paraphrasing even when strings don't exactly match
  - similarity thresholds below 10%-15% are considered ordinary in non-plagiarized work (Mahian et al. 2017)

→ 5% overall similarity between machine answers and the comparison corpora

→ none of the passages were marked as similar to the training corpus we used in fine-tuning

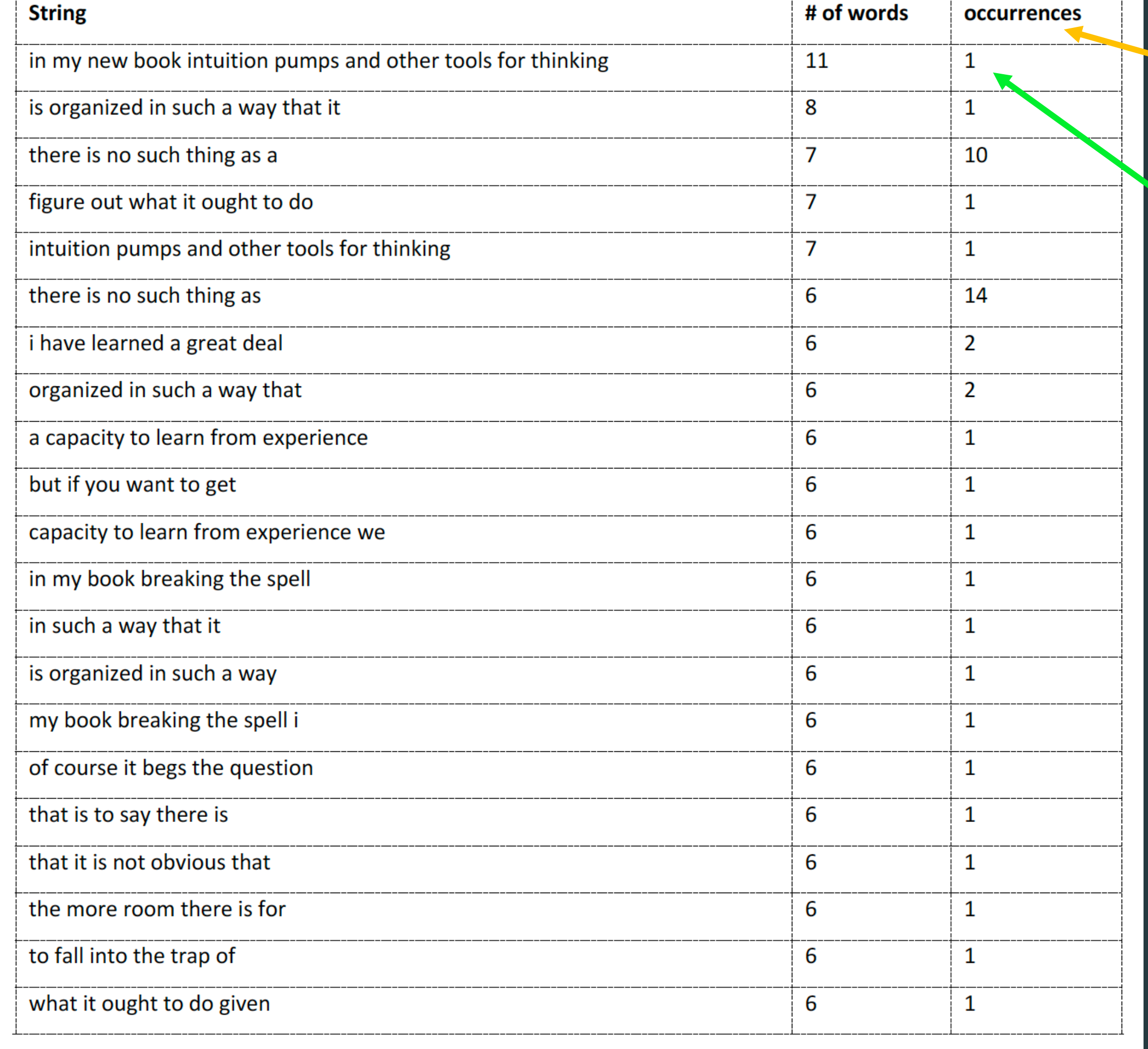
BUT Turnitin plagiarism check process is non-transparent → more transparent process of searching for matching strings

ngram package from the R programming language (Schmidt & Heckendorf 2015)

- looked for strings of 6 or more words that matched between the 3240 words of machine answers & approximately two million words of Dennett's corpus
  - strings defined as contiguous "6-grams," "7-grams," etc., with matching cases sharing the same order of six (or more) words
  - n-grams that appeared exclusively as a subset of a larger n-gram were excluded

WE FOUND 21 MATCHING STRINGS

21 n-gramms found



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

Strings of six or more words that match between the GPT-3 outputs and the Dennett training corpus

number of separate training data segments in the training corpus in which that phrase appears
* occurrences total for shorter strings excludes occurrences in larger matching strings

"intuition pumps and other tools for thinking" occurs once outside of "in my new book intuition pumps and other tools for thinking."

- most of the overlap is in stock phrases of the type favored by analytic philosophers
  - "it is not obvious that" / "in such a way that it" / few book titles
- 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 3 times in 2 machine answers as well as in 26 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.

Our machine is not simply "plagiarizing" Dennett
but rather generating conceptually novel (even if stylistically similar) content!

Fun fact

Our machine might in some respects be a **“supernormal” Dennett** – even more prone to fall into Dennett’s favorite patterns of phrasing than Dennett himself is.

we ran the same ngram check on Dennett’s answers (comprising 747 words)

- taking into account that Dennett’s answers are in total only about one-fourth the length this constitutes less match to the corpus
    - 1 *nine-word string* “exactly what the frogs eye tells the frogs brain” (one occurrence in the corpus)
        - 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
    - 1 *seven-word match* “has a lot to do with the”
    - 1 *six-word match* “life is nasty brutish and short” (a famous quote from Hobbes).
    - *five-word strings:* 72 occurrences in the training data of 18 different 5-word strings in Dennett’s answers

Conclusions

OUR FINE-TUNED GPT-3 CAN SOMETIMES GIVE OUTPUTS INDISTINGUISHABLE FROM DENNETT'S OUTPUTS

- *ordinary research participants untrained in philosophy: near chance / experts on Dennett's work: half of the time*

BUT neural networks are not reliable; they're not like calculators, which always generate the same correct answer.

EXPERT RESPONDENTS PERFORM BETTER THAN NON-EXPERT RESPONDENTS

EXPERT RESPONDENTS WILL RATE DENNETT'S ACTUAL ANSWERS AS MORE DENNETT-LIKE THAN GPT-3'S ANSWERS

✓

EXPERT RESPONDENTS ON AVERAGE GUESS CORRECTLY AT LEAST 80% OF THE TIME

✗

❖This is not a Turing test.

- No back-and-forth.
- We assume that in a proper Turing test, Dennett experts would have reliably distinguished Dennett from our language model.

What now?

ONE COULD THINK ABOUT PRODUCING LONGER OUTPUTS

- evaluated outputs were relatively short (ranging from 38 to 147 words) and thus lacked much argumentative structure

ONE COULD THINK ABOUT TRAINING THE LARGER ENGINE WITH ERIC'S BLOG

- output resembled an extended philosophical argument (448 words long) containing two novel thought experiments and substantial argumentative structure BUT the argument lacked philosophical merit

BUT

before aiming for new large language models accompanied risks should be evaluated

Copyright

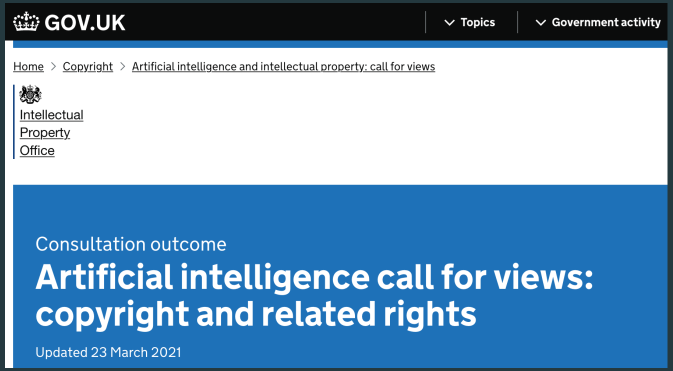
Copyright law governing fine-tuned language models is not yet settled

- unclear whether it is fair use of intellectual property to fine-tune a language model on the works of a single author
    - idea-borrowing via fine-tuned language models might be undetectable as plagiarism, even if it is rightly considered plagiarism
    - fine-tuned models will not output a long sequence of text that exactly matches a sequence of text from the author's corpus

- until the law is settled

WE RECOMMEND SEEKING THE EXPLICIT PERMISSION OF THE AUTHOR BEFORE FINE-TUNING & PUBLISHING ANY OF THE OUTPUTS

- open question
    - How to deal with works by deceased authors? (Nakagawa & Orita 2022)


GOV.UK
Topics
Government activity
Home
Copyright
Artificial intelligence and intellectual property: call for views
Intellectual Property Office

Consultation outcome

Artificial intelligence call for views: copyright and related rights

Updated 23 March 2021

Overreliance & counterfeiting

NOT GOOD ENOUGH!

Dennett-tuned version of GPT-3 did not reliably produce outputs representing Dennett's views.

- not surprising: all deep learning networks do have problems with reliability

(Alshemali & Kalita 2020; Bosio et al. 2019)

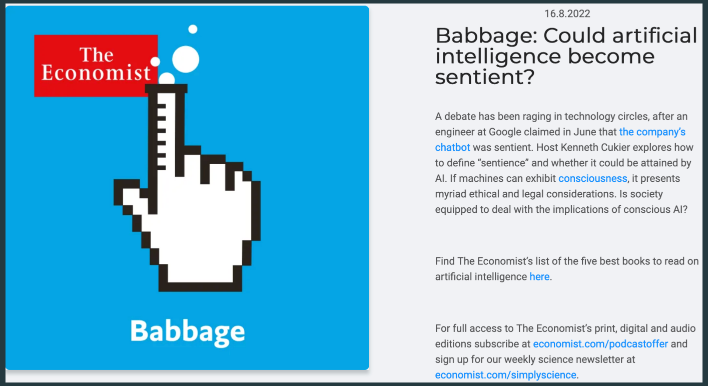
→ user might mistakenly assume that outputs are likely to reflect the actual views of the author

- tempting for students, social media users, or others who might rather query a fine-tuned model of an author than read the author's work

WE RECOMMEND SUBSTANTIAL CAUTION
BEFORE RELEASING TO THE PUBLIC ANY LANGUAGE MODELS FINE-TUNED ON AN INDIVIDUAL AUTHOR

Language models 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



16.8.2022

Babbage: Could artificial intelligence become sentient?

A debate has been raging in technology circles, after an engineer at Google claimed in June that the company's chatbot was sentient. Host Kenneth Cukier explores how to define "sentience" and whether it could be attained by AI. If machines can exhibit consciousness, it presents myriad ethical and legal considerations. Is society equipped to deal with the implications of conscious AI?

Find The Economist's list of the five best books to read on artificial intelligence here.

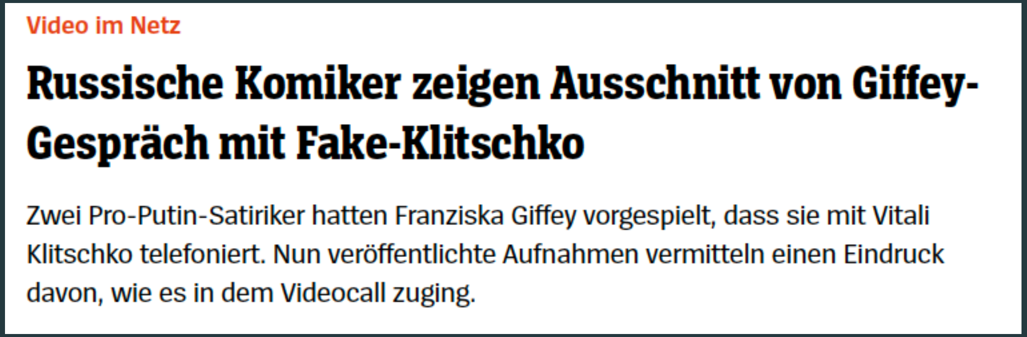
For full access to The Economist's print, digital and audio editions subscribe at economist.com/podcastoffer and sign up for our weekly science newsletter at economist.com/simplyscience.

Increasingly difficult to distinguish

How can teachers in the future ensure that submitted essays are not simply a product of a language model?

- Perhaps universities will return to supervised essay writing in person.

- How can we know whether in chat conversations we are interacting with humans and not chat-bots?



Video im Netz

Russische Komiker zeigen Ausschnitt von Giffey-Gespräch mit Fake-Klitschko

Zwei Pro-Putin-Satiriker hatten Franziska Giffey vorgespielt, dass sie mit Vitali Klitschko telefoniert. Nun veröffentlichte Aufnahmen vermitteln einen Eindruck davon, wie es in dem Videocall zuging.

How can we avoid new forms of plagiarism?

- How to deal with verifiable authorship with respect to the mass of electronically distributed texts?
- Will we establish new social practices that aim at proving that one is really the original author of what is written?

Long-term potentials

IS PHILOSOPHY SAFE FROM AI TAKEOVER?

What do you think about
- computer programs that generate music in the style of a particular composer
- image-generation programs like OpenAI's Dall-E
- language-models that generate text on behalf of the user

LANGUAGE MODEL as THINKING TOOL USED BY HUMANS

- future fine-tuned language models might produce outputs interesting enough to serve as a valuable source of cherry-picking by experts
- selected outputs might have substantial merit

→ an author might create many outputs, choose the most promising, edit them lightly, and present them, not unreasonably, as original work

Long-term potentials

June 2022:
Google engineer Blake Lemoine said he became convinced that LaMDA was sentient

SENTIENT, CONSCIOUS, COMPREHENDING?

Our GPT-3 language model of Dennett does not have Dennettian philosophical opinions about consciousness, God, and animal suffering.

- As long as language models of philosophers remain focused on next-word prediction, apparently with no cognitive model of the world, it is unlikely that they have philosophical beliefs.

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, or insight, or common sense, potentially triggering new philosophical ideas in the reader, and perhaps also paving the way for the creation of artificial entities, eventually, who are genuinely capable of philosophical thought.

Take home message

Our fine-tuned GPT-3 can sometimes give outputs indistinguishable by experts from Dennett’s outputs.

- BUT neural networks are not reliable; they’re not like calculators which always generate the same correct answer.

Fine-tuned language models can create opportunities for plagiarism, over-interpretation, and over-reliance

- Our efforts to make sense of anything that looks roughly interpretable can betray us!
- GPT-3 can serve as an automatic plagiarist → dangerous prospect of this technology because copyright doesn’t come close to dealing with it!

RECOMMENDATIONS

- We need legislation to outlaw some of the ways in which these systems might be used!
- We should always ask for allowance if we build a model based on a living person!

❖ But it could also be a helpful thinking tool !

Acknowledgements

This could have not happened without Eric and David Schwitzgebel!

Special thanks to both Daniel C. Dennett & 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)

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-networkmodels/
- 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-becomesentient
- 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-songsraise-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-computerscientists-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

REFERENCES

- Fawzi, A. et al. (2022). Discovering novel algorithms with AlphaTensor. https://www.deepmind.com/blog/discovering-novel-algorithms-withalphatensor?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-forviews/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-learninglanguage-generator-gpt-3-nlp

- 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-learninglanguage-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-arenot-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

REFERENCES

- 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-wrongquestion/
- 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-thinkrobots-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

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 $p < .001$).

https://osf.io/vu3jk/?view_only=2970a846490842a48919e15f6aa0a6cc