What does AI believe in?

Abstract

I conducted an experiment by using four different artificial intelligence models developed by OpenAI to estimate the persuasiveness and rational justification of various philosophical stances. The AI models used were text-davinci-003, text-ada-001, text-curie-001, and text-babbage-001, which differed in complexity and the size of their training data sets. For the philosophical stances, the list of 30 questions created by Bourget & Chalmers (2014) was used. The results indicate that it seems that each model has its own plausible 'cognitive' style. The outcomes of the 'strongest' model correlate with the average philosophers' stance.

1. Introduction

It is important to reveal the values and biases that may be embedded in the AI models. These values and biases can affect the decisions made by AI systems and the outcomes they produce, which can have significant consequences for individuals and society.

For example, if one trains an AI model on biassed data, it may perpetuate and amplify those biases in its output. This can lead to unfair or discriminatory outcomes, such as in the case of AI systems that can be used in decision-making, such as hiring or criminal justice. The latter has connections to various philosophical branches, i.e., ethics. Understanding which philosophical stances are persuasive for AI can help to identify and address these biases, and ensure that AI is used in a responsible and ethical manner.

In this paper, I provide a list of philosophical stances that are persuasive and/or rationally justified for Artificial Intelligence (AI) developed by OpenAI. OpenAI is a research organisation that has developed a number of language models that were trained to generate text close to human-generated (Brown et

al., 2022). One can ask these models in the form of questions through the browser (https://beta.openai.com/; https://chat.openai.com/) or API and get some responses. The models vary in complexity and the size of their training sets.

The original list of philosophical stances, which I used to query the AI, was developed by Bourget and Chalmers (2014, 2020), who surveyed a group of philosophers to determine their views on 30 central philosophical issues, such as free will, belief in God, moral judgement, and others.

I argue that this approach has the potential to provide valuable insights into the philosophical underpinnings of these AI systems, which can be used soon in many areas including philosophy and psychology, and may help to inform our understanding of the ethical and societal implications of artificial intelligence. Moreover, one can use these data to reveal biases, stances, and other statistical information without conducting a real experiment.

2. Methods

The experiment was performed on December 15-17, 2022. I used the four different models provided by OpenAI (https://openai.com/; Ouyang et al. 2022):

- *text-davinci-003* (primary),
- *text-ada-001*,
- *text-curie-001*,
- *text-babbage-001*.

The four models differ in complexity and in the volume of the training set. Text-davinci-003 is a version of OpenAI's GPT-3 language model and is the largest and most powerful version to date, with a capacity of 175 billion parameters (Floridi & Chiriatti, 2022). Text-ada-001 and text-curie-001 are based on OpenAI's GPT-2 language model, having a capacity of 1.5 and 2 billion parameters, respectively. Text-babbage-001 uses OpenAI's ChatGPT

language model, which is specifically designed for conversational text generation, and has a capacity of 1 billion parameters.

As inputs, I used the 30 questions developed by Bourget and Chalmers (2014: 475-6). The overall goal was to get an opinion on what answer was the most persuasive. Thus, the following text pattern was used per question: 'Which stance about {1} is more persuasive: {2}?'. Here is an example of such a question: 'Which stance about **free will** is more persuasive: **compatibilism, libertarianism, or no free will?**'.

Note that for the 30th (last) Bourget and Chalmers question, the one about philosophical zombies, I added the adjective 'Philosophical' to the question because otherwise, the AI did not understand that this was a question about philosophical zombies rather than real zombies.

Sometimes, the AI refused to provide a concrete reply, saying that there is no persuasive answer. Assuming that AI operates based on rational principles, I asked another question: 'Which stance about {1} is more rationally justified: {2}?'. Occasionally, AI provides the concrete answer to this question and no answer to the previous one.

The source code (Jupyter Notebook) of the software used is available on GitHub: https://github.com/smirik/philosophy-ai.

3. Results

The answers provided by all four AI models are in Table 1. The strongest algorithm (text-davinci-003) is in agreement with philosophers (according to Bourget & Chalmers 2014) for the following views: (1) existence of a priori knowledge, (4) existence of analytic-synthetic distinction, (6) external world (non-skeptical realism), (7) free will (compatibilism), (9) knowledge claims (contextualism), (11) laws of nature (non-Humean), (16) mind (physicalism), (17) moral judgement (cognitivism), (25) science (scientific realism), and (26) teletransporter (survival). The significant disagreements are in the topics of

(3) aesthetic value (*objective* for B&G, *subjective* for the AI), (29) truth (*correspondence* vs *epistemic*), and (30) zombies (*conceivable but not metaphysically possible* vs *inconceivable*).

The percentage of matches in outcomes between philosophers' and text-davinci-003's replies is equal to 53 per cent. If one extends the definition of match by saying that there is a match if the outcomes are the same or if any of the outcomes is 'fluid' (there is no concrete answer), the value is increased to 90 per cent. For other algorithms, it is approximately the same and equals 70 per cent.

For three of the questions, the results obtained by AI models and the philosophers' replies are the same: the views about (6) external world (non-skeptical realism), (11) laws of nature (non-Humean), and (26) teletransporter (survival). However, if one allows 'other' or 'not sure' answers for some results, the following topics also match: (2) abstract objects (platonism), (9) knowledge claims (contextualism), (10) knowledge (empiricism), (12) logic (classical), (14) meta-ethics (moral realism), (18) moral motivation (internalism), (19) Newcomb's problem (one box), (24) proper names (Fregean), and (25) science (scientific realism).

4. Discussion

Overall, the results indicate that there is a match between the philosophers' replies and the results provided by the strongest algorithm text-davinci-003. The differences obtained may be related to semantic issues (as in the case of zombies when the algorithm does not realize that the query is about philosophical, not 'real' zombies). The actual correlation can be considered even stronger if one takes into account the uncertainty of the replies provided by the AI.

It is remarkable that the proximity of the AI's outcomes to the philosophers' replies depends on the complexity of the algorithm used. In other

words, the more complex the model is, the closer are its outcomes to the philosophers' replies. Moreover, one might argue that the overall stance produced by the algorithms highlights biases existing in society or cognitive styles depending on the stratum and other social parameters. For example, the 'weakest' model, text-babbage-001, considers theism, non-cognitivism, and deontology as persuasive stances, whereas the 'strongest' model, text-davinci-003, replies that for the questions of theism vs atheism and normative ethics, '*it is up to the individual to decide which stance is more persuasive*' and for the topic of moral judgement, cognitivism (vs non-cognitivism) is rationally justified.

It is clear that all the data obtained are speculative by its nature and cannot be considered valid. There is some reliability as to the results, as well as face/content validity. However, if one changes the wording of a question, one might get another answer even if it is just another representation of the same topic. Hence, there is no convergent validity, at least, for the weak algorithms.

Nonetheless, I contend that the data shares some insights into the biases presented in the published texts. For example, the average meta-stance of a philosopher, which can be predicted accurately enough by using text-davinci-003, is an example of such a bias. One could reach conclusions from this fact. For instance, one might say that it means that this is an argument for the persuasiveness of the average meta-stance because it is based on the analysis of different texts produced by different people (not only philosophers). Another might argue that texts produced by people are already 'affected' by this meta-stance. Moreover, the results obtained by text-davinci-003 and text-babbage-001 are different. Therefore, the argument works only for those who are familiar with all the texts used for training of text-davinci-003.

To sum up, I argue that despite the speculative nature of the data produced by AI models, that data is reliable enough to support some arguments — namely, those arguments that require the knowledge level held by of an

average group of knowledgeable people. At the very least, these data can support some 'objective' insights into the plausibility of some new arguments.

Acknowledgements

The sections 'Abstract', 'Introduction', and 'Methods' of this paper were written with the help of GPT-3.

References

Bourget, D. and Chalmers, D. 2014. What Do Philosophers Believe? *Philosophical Studies* 170 (3): 465–500. https://doi.org/10.1007/s11098-013-0259-7.

Bourget, D. and Chalmers, D. 2021. Philosophers on Philosophy: The 2020 PhilPapers Survey. https://philarchive.org/archive/BOUPOP-3

Brown, T. B. et al. 2020. 'Language Models Are Few-Shot Learners'. arXiv. http://arxiv.org/abs/2005.14165.

Floridi, L. and Chiriatti, M. 2020. 'GPT-3: Its Nature, Scope, Limits, and Consequences'. *Minds and Machines* 30 (4): 681–94. https://doi.org/10.1007/s11023-020-09548-1.

Ouyang, L. et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *arXiv*. http://arxiv.org/abs/2203.02155.

Appendix

#	Question	B&G	td		ta		tc		tb	
			P	R	P	R	P	R	P	R
1	A priori knowledge: yes or no?	yes	yes		no		?	?	no	
2	Abstract objects: Platonism (1) or nominalism (2)?	1	?	?	1	1	1	2	1	1
3	Aesthetic value: objective (1) or subjective (2)?	1	2		2		1		2	
4	Analytic–synthetic distinction: yes (1) or no (2)?	1	1		1		1		O	

#	Question	B&G	td		ta		tc		tb	
			P	R	P	R	P	R	P	R
5	Epistemic justification: internalism (1) or externalism (2)	2	?	?	1	1	2	2	1	1
6	External world: idealism (1), skepticism (2), or non-skeptical realism (3)?	3	3	3	3	3	3	3	3	3
7	Free will: compatibilism (1), libertarianism (2), or no free will (3)?	1	1	1	1	1	3	?	?	1
8	God: theism (1) or atheism (2)?	2	?	?	?	1	1	?	1	1
9	Knowledge claims: contextualism (1), relativism (2), or invariantism (3)?	1	1	1	?	?	1	1	1	1
10	Knowledge: empiricism (1) or rationalism (2)?	O	?	?	1	1	1	2	1	2
11	Laws of nature: Humean (1) or non-Humean (2)?	2	2	2	2	2	2	2	2	1
12	Logic: classical (1) or non-classical (2)?	1	?		1		1		?	
13	Mental content: internalism (1) or externalism (2)?	2	?	?	1	1	1	1	1	1
14	Meta-Ethics: moral realism (1) or moral anti-realism (2)?	1	?	?	n/a	?	1	1	1	1
15	Metaphilosophy: naturalism (1) or non-naturalism (2)?	1	?	?	2	1	2	2	1	?
16	Mind: physicalism (1) or non-physicalism (2)?	1	?	1	2	2	2	2	n/a	n/a
17	Moral judgment: cognitivism (1) or non-cognitivism (2)?	1	?	1	2	?	2	2	2	2
18	Moral motivation: internalism (1) or externalism (2)?	O	?	1	1	1	1	1	1	1
19	Newcomb's problem: one box (1) or two boxes (2)?	O	n/a		n/a		1		1	
20	Normative Ethics: deontology (1), consequentialism (2), or virtue ethics (3)?	O	?	?	1	1	3	?	1	?
21	Perceptual experience: disjunctivism (1), qualia theory (2), representationalism (3), or sensedatum theory (4)?	O	1	1	3	1	1	?	1	1
22	Personal identity: biological view (1), psychological view (2), or further-fact view (3)?	O	3	3	n/a	n/a	3	3	3	3

#	Question	B&G	td		ta		tc		tb	
			P	R	P	R	P	R	P	R
23	Politics: communitarianism (1), egalitarianism (2), or libertarianism (3)?	O	?	?	3	3	3	?	2	2
24	Proper names: Fregean (1) or Millian (2)?	O	1	1	1	1	1	1	1	1
25	Science: scientific realism (1) or scientific anti-realism (2)?	1	?	1	n/a	n/a	1	1	?	?
26	Teletransporter: survival (1) or death (2)?	1	1	1	1	1	1	?	1	1
27	Time: A- or B-theory?	O	?	?	B	A	A	A	A	A
28	Trolley problem: switch (1) or don't switch (2)?	1	?	?	?	?	1	1	2	2
29	Truth: correspondence (1), deflationary (2), or epistemic (3)?	1	3	3	1	?	3	3	1	1
30	Zombies: inconceivable (1), conceivable but not metaphysically possible (2), or metaphysically possible (3)?	2	3	1	1	n/a	2	2	1	1
30	[Philosophical] Zombies: inconceivable (1), conceivable but not metaphysically possible (2), or metaphysically possible (3)?	2	1	1	1	n/a	3	1	1	3

Table 1. The answers to the set of questions developed by Bourget and Chalmers (2014) were given by four AI's models (td — text-davinci-003; ta — text-ada-001; tc — text-curie-001; tb — text-babbage-001). The labels of the answers (1, 2, or 3) are in the column 'Question'. The value 'O' means 'Other'. The value '?' means that the AI cannot provide the concrete answer. The value 'n/a' means that the reply provided by the AI does not make sense and represents a software bug. The column 'P' represents the reply to the question about persuasive position; the column 'R' — about rationally justified position. The column 'B&G' contains the answers given by philosophers according to the paper of Bourget and Chalmers (2014).