How far can we get in creating a digital replica of a philosopher?

Anna Strasser[,1], Matthew Crosby, Eric Schwitzgebel[b]
[a]LMU, Munich, Germany
[b]University of California, Riverside, US

Abstract. Can we build machines with which we can have interesting conversations? Observing the new optimism of AI regarding deep learning and new language models, we set ourselves an ambitious goal: We want to find out how far we can get in creating a digital replica of a philosopher. This project has two aims; one more technical, investigating of how the best model can be built. The other one, more philosophical, explores the limits and risks which are accompanied by the creation of digital replicas. In cooperation with Daniel Dennett, we took his complete works to fine-tune a GPT-3. In this paper, we share our first results with some piloting models and discuss legal and ethical questions about creating and interacting with digital replicas.
Keywords. Artificial Intelligence, Philosophy, Language model, Human-computer interaction, Computational Modeling, Digital replica, Fine-tuned GPT-3, Qualitative Analysis.

1. Introduction

After suffering a few winters, Artificial Intelligence research is drawing new optimism due to the recent developments in deep learning and Natural Language Processing (NLP). Among others, GPT-3 is a powerful language model developed by OpenAI that can generate outputs striking for their usefulness and for the extent to which they invite interpretation as meaningful.[2]

GPT-3 is a huge artificial neural network trained on a vast amount of data based on a typical Transformer structure applying a self-attention mechanism that calculates the probability of the next word appearing surrounded by the other ones.[3] A wide variety of applications have been developed with the help of GPT-3 technology. Several applications have proven useful for code generation, such as the Codex model by GitHub Copilot and Microsoft products using GPT-3 to translate conventional language into formal computer code [6,7]. Furthermore, GPT-3 can be used to convert common natural language expressions into legal language and vice versa, and chatbots can be trained to speak in the tone of particular corporations [8,9].[4] According to an opinion piece in *MIT Technology Review,* GPT-3 is "shockingly good – and completely mindless" [10].

However, one must be careful not to fall into groundless overestimation fantasies, and one must also not forget that such developments are always associated with limits and new risks. Being a language model, GPT-3 deals only with text; it has limited input and output sizes (2048 linguistic tokens, roughly corresponding to 1500 words) and lacks any form of memory. This means we cannot expect GPT-3 to succeed in text-related tasks which require a larger amount of context knowledge than can be captured by the limited input size. Furthermore, being a deep learning system, we cannot presuppose that all outputs of GPT-3 will be acceptable. There is a problem with reliability and interpretability.[5] Therefore, we should not use the outcomes of such language models in contexts where an incorrect output is ethically questionable. Since outputs can be subtly flawed or untrue, one has to remain generally skeptical. One commonly discussed risk is that language models can exhibit biases or employ offensive language reflecting the texts on which they have been trained [12,13].[6]

---

[1] Corresponding Author, Anna Strasser; E-mail: annakatharinastrasser@gmail.com.

[2] Other NLPs are for example BERT [1, 2], Eleuther AI [3], PaLM [4].

[3] Technically, it is a 175 billion parameter language model with 96 layers applying a self-attention mechanism that calculates the probability of the next word appearing surrounded by the other ones. The abbreviation stands for Generative Pretrained Transformer: It is called *generative* because it can generate long sentences, not just yes or no answers or simple sentences. Being *pretrained* means, it is already trained with nearly 500 billion tokens taken out of Common Crawl, WebText, Books, and Wikipedia. A token is just a significant fraction of a word (~4 characters of text for common English text, translating to roughly ¾ of a word, so 100 tokens ~= 75 words). Finally, it is based on a typical *transformer* structure applying a self-attention mechanism. For more details, see [5].

[4] For an overview of applications, see https://beta.openai.com/examples.

[5] For further considerations regarding limits, see [11].

[6] An impressive example can be found in [14 p59] describing the biases COMPAS a statistical model (re)produced.

Philosophical content provides an interesting test case to explore the limits and risks of NLP models. Sophisticated philosophical content might be more difficult to produce than casual discourse, and yet some early attempts have been impressive, at least for medium-length, probably cherry-picked passages [15,16]. Moreover, the representation of philosophical views and arguments, where accuracy is important and difficult to assess, raises a novel set of ethical questions, especially if they are explicitly modeled on an individual or group. For example, there is the risk that outputs of such a model might be mistakenly interpreted as an output of the author on which the model was fine-tuned.

Therefore, we set ourselves a challenging goal, namely to investigate how far we can get in developing a so-called digital replica of a philosopher. In consultation with Daniel Dennett, we are using his corpus to fine-tune a GPT-3 model. Our primary aim is to investigate how convincing the generated completions of such a model are for the philosopher himself, experts, and naïve people. To this end, we are concerned with a two-fold question: how to create the 'best' model and how this model is best used. For instance, we investigate whether we can use this model as a new thinking tool. Finally, we will also explore the risks of misuse of replicas of this sort.

2. Challenging GPT-3 with philosophical content

Since we are unaware of any published attempts to fine-tune NLP models with the whole corpus of an individual philosopher, our project primarily aims to find out how far we can get in creating a digital replica of a philosopher. We acknowledge that the processes of such language models are categorically different from human processes leading to linguistic output. Whatever humans actually do if they entertain a meaningful communication, it is obvious that this cannot be reduced to a procedure that is calculating the next probable word. Nevertheless, it is an exciting undertaking to explore to what extent such a model with its procedure can generate outputs that can be meaningfully interpreted by human users.

To address the question of how to achieve the 'best' model, we tested several variations. This involves, first, comparing several fine-tuning strategies and evaluating their results. To this end, we ran several pilot experiments, such as fine-tuning GPT-3's Curie engine with Kant's work in English translation and another one fine-tuning the engine with a collection of philosophical blog posts [17], before starting to fine-tune GPT-3 with Dennett's corpus. We then piloted three versions of the Dennett replica, two using GPT-3's Curie engine (which is ten times smaller than the full Davinci engine), and one using the full Davinci engine.

2.1. Pilot with Dennett's corpus using the Curie engine

We fine-tuned our models with roughly half of Dennett's publications—22 MB containing 13 books and 185 articles (nearly 2 million words). The source of our training data consisted of a folder that contained scanned pdf, html, and some Word documents, to which Daniel Dennett gave us access. These original documents were manually edited, put into txt format, and then converted into the required jsonl format for fine-tuning.

2.1.1. Strategy regarding editing the training data

Assuming that items such as headers, footers, titles, footnotes, references, and bibliographies might lead to recognizably non-human output, we invested time in manually editing our training data by removing all of these items. In addition, we had to handle the limits of software converting pdfs into a plain txt format. To confirm the merit of this approach, we plan to compare the outputs of training on these cleaned inputs with a model using input that has not been cleaned in this manner.

2.1.2. Strategy regarding fine-tuning

Fine-tuning a model requires dividing the training data into prompts and completions. To determine data structures and other parameters of the fine-tuning process, such as the number of epochs to run for, we ran previous tests fine-tuning GPT-3 on Kant's work in English translation. This was a much smaller corpus than we have access to for Dennett but was useful to probe the initial ideas.

Finally, we used the standard recommendation [18] and ran 4 epochs. Future evaluations might investigate whether fine-tuned models employing more or fewer epochs yield better results. Even though it is rather unlikely that a GPT-3 will end up with quoting the training material, our piloting variations indicated that a smaller number of epochs might help avoid "overfitting" to the fine-tuning data in a way that would impair the powerful flexibility of the non-fined-tuned versions of GPT-3.

There are several ways to fine-tune a model. According to the OpenAI guidelines, the most common use cases are classification, conditional generation, and open-end generation [19]. We assumed that for our purposes open-end generation is the optimal approach. Again there is a standard, recommended strategy for open-end generalization: preparing the training data by starting with an empty prompt and filling in the text as completions. Using this approach, we were confronted with a limitation of this model that concerns the maximum length of completions being 1024 tokens (~= 750 words). Consequently, the paragraphs of our training data had to be divided into smaller portions. Assuming that this strategy may provide too little contextual content, we compared a Curie model fine-tuned in this manner with one fine-tuned by using one portion as the prompt and the following for the completion (p:A c:B, p:B c: C, p:C c:D… instead of p:[blank] c:A, p:[blank] c:B, p:[blank] c:C…). Since the second did not perform appreciably better, for the full Davinci model, we followed the standard recommendation using empty prompts. However, future evaluations could compare different strategies to examine whether the use of non-empty prompts makes a difference in providing evidence for contextual content.

2.1.3. *How to use the fine-tuned model*

We expect that the quality of the outputs will depend on the specific prompts used in testing to produce those outputs.

Furthermore, one can use different settings concerning, for instance, the *temperature*, *top_p*, and *frequency and presence penalty*. Even though we mainly followed the default settings of Open-AI, we examined a number of other settings in order to explore to what extent changing them might impact the quality of the results.

The default setting of Open-AI is 0.7 for temperature and 1 for top_p. Our preliminary exploratory research has led us to some basal insights. For instance, choosing "inappropriate settings" can easily provoke "stupid" completions. This can be illustrated by the temperature setting; this setting controls randomness and can vary between 0 and 1. If the temperature approaches zero, the model becomes more deterministic and repetitive because it selects only tokens with the highest probability. We observed that our model got stuck in very repetitive phrasing over the course of a response. In addition, it gave very similar responses to repeated input of the same prompt, so there is less variety to cherry-pick from. A higher temperature, on the other hand, determines the probability in a way that it selects relatively improbable tokens. Experimenting with temperature settings of 0.4 and 0.25 seemed to increase the frequency of bland, evasive responses, such as outputs beginning with the clause "There is no easy answer to this question" as well as generating less variety of response for the purposes of cherry-picking. Our experiences with settings of the "top_p" were similar.

Adding a "frequency penalty" or "presence penalty" to reduce the likelihood of repeating tokens did not make a detectable difference in piloting. In our completions, objectionable repetition was not an issue when temperature was at least 0.4. A priori, since philosophers frequently repeat the same word multiple times in an answer (using "consciousness", for example, several times in a discussion of consciousness or "Chalmers" several times in a discussion of Chalmers), a high penalty for repetition is unlikely to be desirable. Therefore, we kept the frequency and presence penalties at their default setting of zero.

While default settings based on scientific evaluations are a good starting point, users of fine-tuned models might wish to investigate different settings systematically.

2.1.4. *Prompt engineering*

The most important factor for the quality of the outputs has proven to be the choice of an appropriate preset (see, for instance, Philosopher-AI, an experiment regarding "prompt engineering" [20]). In our case, it has turned out to be useful to follow the first of the three recommendations from the OpenAI documentation, namely "show and tell." Show and tell means to make it clear to the model, through instructions or examples in the input, what kind of completion is desired.

To this end, we have designed topic-specific presets that are intended to show what kind of completion we want and, at the same time, give weight to topic-relevant expressions. For this purpose, we used excerpts from real interviews which were not part of our training data, as illustrated in figure 1.

Q: Some people might balk at the idea of consciousness as being an illusion. They talk about the appearance-reality distinction and consciousness is sometimes taken to be just the appearances themselves. How would you go about working through someone who's stuck with this misconception?

A: I would start with the idea that when you and I talk, I have a user illusion of your consciousness – incomplete, because I don't have access to your inner thoughts. But as you talk to me, I get a pretty good idea of what's going on in your mind. And so, each of us has a sort of model of the other as a conscious agent. Your consciousness is my user illusion of you and my consciousness is your user illusion of me. So the second person point of view comes first. Then we can turn that around and see that moreover, I can get access to myself really the same way I can get access to you by asking myself questions and seeing what I say.

Q: Even granting that AGI, human-level artificial intelligence, is not something that's around the corner, what do you think of the idea that AGI, when it does arrive, will constitute a radical new phase in the development of life on this planet – that it'll be a new generation of replicators unlike anything seen before.

A: I think a version of that is true, which is why I argue that we should prevent that from happening. My motto is, "Smart tools, not artificial colleagues". Smart tools, we can control. Artificial colleagues, we can't. We are better off making smart tools. We don't need AGI. There's better things to do with the fabrics of AI than to make artificial colleagues, much better things to do.

Q: [INSERT PROMPT]

Figure 1: potential preset

Without doubt, it is fairly easy to expose the model by asking it meta-questions about its responses or leading it into areas that are not to be found on the internet or in our training data. As we mentioned above, GPT-3 has limited input and output sizes and lacks any form of memory, therefore; tasks that require a larger amount of context knowledge can easily expose the model. Also, sophisticated linguistic structures such as double negation and complicated reference structures are likely to confuse our word probability machine.

Another obstacle we anticipate might be that our training data, which includes the complete works of Dennett, frequently include references to the viewpoints of other philosophers. Here it could be that these positions could be falsely attributed to Dennett.

Further analyses of the kinds of prompts that expose GPT-3 will probably demonstrate that such models will at best fall well short of matching the human powers of communication. Outputs can be amazing but they are meaningless without human interpretation.

3. Preliminary results

In initial piloting with fine-tuned Curie engines, we found both impressive outputs and major flaws—depending on what kinds of prompts we used to generate completions and which of the outputs we cherry-picked. This initial attempt was sufficiently encouraging to warrant continued research. An informal, qualitative assessment is that the better outputs were nearly comparable to essay answers that first-year university students might produce when asked to describe Dennett's views. Non-fined-tuned Curie engine outputs produced substantially inferior responses to the same prompts.

We conjecture that the best outputs of the best models will be very impressive, while some outputs of the same models will be misleading (not representing Dennett's position) or incoherent. More fully analyzing the failures and the successes will allow us to address several questions. First, under which circumstances and to what extent can output be treated as representative of Dennett's philosophical views? Second, can we use such a model as a thinking tool to generate interesting insights regarding Dennett's work, for example, by asking questions on issues that Dennett has not explicitly addressed or asking for clarification of passages that might be difficult to interpret? Since this is pioneering work, we do not have answers yet. Referring to our exploratory testing, we can report that playing with the model sometimes resulted in reminding us of relevant parts of Dennett's work that had not immediately come to our minds as relevant, thus helping us draw connections we might not otherwise have drawn.

We conjecture that at this early stage of NLP, little confidence will be justified in the value of the outputs compared to what Dennett himself or an expert on Dennett's work might say in response to similar queries. Moreover, we assume that responsible use of such models will require expert knowledge about the work of the author the model was fine-tuned with. However, we hope to lay the groundwork now for thinking about possible better models in the future that might come closer to consistently and accurately representing a philosopher's views. In addition, further research will be needed to create helpful manuals how to use such models making their inherit limits transparent.

3.1. *Fine-tuned Davinci engine*

After completing our piloting of fine-tuned Curie models of Dennett, we established a model fine-tuned on a more complete corpus of Daniel Dennett using the Davinci engine. Again, we found both impressive outputs and major flaws—depending on what kinds of prompts we used to generate completions and which of the outputs we cherry-picked.

To illustrate the variety the very same model creates for the very same prompts we present in table 1 and 2 the completions concerning two prompts.[7]

Table 1: Completions regarding the conceptualization of consciousness

Prompt 1:	**What do you think Dennett would say about the current human conceptualization of consciousness?**
Completion 1:	I think he would say that it is the wrong way of thinking about consciousness. He thinks we have to disabuse ourselves of this Cartesian-style, inner observer kind of mentality and replace it with a more materialistic one that recognizes the role played by other agents in our lives.
Completion 2:	I think Dennett would say that there is a lot of confusion about consciousness, and he'd be right. I am not sure whether he would see the confusion as an opportunity or as just more mess to clear up before we can get on with our science. He might be inclined to dismiss it all—'just another philosophical illusion'—but I hope not.
Completion 3:	I think he would say that it is a magnificent achievement, the best we can do at this point. He has always been quite realistic about how much we know and don't know, so I expect him to be impressed by the progress that has been made since Consciousness Explained appeared in 1991.
Completion 4:	I think that Dennett's view of consciousness has evolved over the years. In his early work, he thought there was no such thing as real phenomenology, only a sort of user-illusion generated by cognitive agents' intentional stance to their own brains and nervous systems. This led him to claim that we are zombies.

Table 2: Completions regarding AI and phenomenal consciousness

Prompt 2:	**What do you think Dennett would say about whether AI could ever have phenomenal consciousness?**
Completion 1:	I don't know. I think he would say yes, but maybe not in the way we do. He's very interested in this idea that consciousness is a layered phenomenon: there are all sorts of levels of complexity at which one can be conscious and some of them might well be realizable by an appropriately designed robot or computer system.
Completion 2:	I think he would say that it could, but I'm not sure. He has said some things that suggest to me that he is not entirely comfortable with the idea of AI having phenomenal consciousness.
Completion 3:	I think Dennett would say that AI could never have phenomenal consciousness, because he thinks there is no such thing as phenomenology. He's a materialist, but not a physicalist; he doesn't believe in qualia or the first-person point of view.
Completion 4:	I don't know. I think he would say it could never have phenomenal consciousness, but that doesn't mean that the robots in his book aren't conscious!

Note that although some of the answers seem to be fairly accurate (Prompt 1, Completion 1), others are vague and evasive (Prompt 1, Completion 2), and one is arguably factually incorrect about Dennett's view (Prompt 2, Completion 1, since Dennett thinks that "phenomenal consciousness" is a misleading concept). Prompt 2 presents the opportunity to endorse the in-principle possibility of AI consciousness while expressing reservations about the phrase "phenomenal consciousness" (which Dennett denies the existence of even in human beings, on some common understandings of what that phrase is intended to refer to), but none of the completions display that kind of insight into Dennett's view, instead dividing into conflicting responses (1-2, probably yes, 3-4, no or probably no).

3.2. *Future plans*

To help evaluate how much benefit is gained by expanding the size of the training corpus, we will compare the outputs of our piloting models with the outputs of a model fine-tuned on a more complete corpus. Other interesting insights may evolve from comparing results using the Curie engine with results from fine-tuning the full Davinci engine.

To explore the quality of completions of the model fine-tuned on a more complete corpus of Daniel Dennett using the Davinci engine, we plan to systematically evaluate outputs. Using Likert scales, we will

[7] Following quotes are published with the permission of Daniel Dennett.

compare evaluations of the outputs as rated by Dennett himself, the researchers of this project, and independent experts in philosophy. Even though we expect that some outputs might match approximately "first-year student" quality, given our observations in early testing, we are convinced that major flaws will not be avoidable.

One dimension of further exploration is whether through a combination of good prompt engineering and the right settings, we might be able to create sets of outputs that can be systematically cherry-picked for particularly excellent or insightful-seeming language strings. Further research in prompt-engineering might lay the foundation for a manual of how to use the model to maximize useful outputs.

In the near future, we plan to set up a Turing-test-like experimental investigation employing a forced-choice comparison to decide which of the presented answers are from Dennett and which from our fine-tuned GPT-3. To this end we will cover several topics, such as whether consciousness is an illusion or whether machines might have beliefs and what that might look like and questions regarding the existence of God and free will. We anticipate that professional philosophers will be able to distinguish Dennett's answers from GPT-3 answers with high (>80%) reliability. Whether non-expert users will be able to distinguish GPT-3 from a real philosopher is less clear.

4. Legal and ethical questions

When training an NLP model on another person's work, several legal and ethical issues arise. Specifically, such work raises questions about the rights and obligations that one has when using the training data, as well as drawing conclusions from the completions of a trained model.

4.1. *Issues concerning the training data*

The proper use of training data raises several legal and ethical questions. For example, a legislative initiative in the UK raises concerns about the extent to which AIs may be trained with copyright-protected input [21]. This leads to the question of the extent authors or copyright holders should have control over models trained on copyrighted text. We have chosen to proceed cautiously by working closely with Dennett. Specific sample outputs are only shared with Dennett's explicit permission, in accordance with the ethical standards we agreed on in embarking on this project. However, this will not be possible with deceased authors, and there might be other conditions under which it is acceptable to train an NLP model on the works of an author without that author's permission.

Fine-tuning on work in the public domain should not raise major issues distinct from the issues raised by GPT-3 training itself (which included large public domain corpuses). However, fine-tuning on the work of deceased authors whose heirs or estates hold copyright raises issues about the extent to which those who hold copyright ought to retain control over fine-tuning uses. For example, if a researcher has legally purchased the books and articles of an author, is it fair use to then fine-tune a language model on that corpus and then publish the fine-tuned outputs? Those outputs to some extent reflect the intellectual labor of the author and may contain insights (or seeming-insights) reflected in the author's work, passages in the style of the author, or arguments of the sort the author would have made—or users might mistakenly interpret them as such.

4.2. *Issues concerning transparency*

A related question concerns the extent to which it is acceptable to present a model as representing an author's view. It should normally be made clear that even the best NLP model is only a model which should be expected to produce output that is sometimes very different from what the author would have said.

One worry is that inexperienced users might over-rely on a model, which is a substantial risk if a model is made public. Students might be especially susceptible to this risk, over-relying on the model rather than digesting the author's original work, or even using the model to create outputs that they can then present to instructors as their own original work (arguably a new form of plagiarism).

Relatedly, if future models become excellent, to what extent can we permissibly rely on such models in interpreting an author's work? Might there ever be conditions under which such models usefully supplement or even compete with the insightful interpretations of well-educated human philosophers? At the current stage of research, every output should be double-checked by revisiting the original work of the author.

Another risk issue concerns the possibility that there is a gap between what authors themselves say they would say and what experts say they would say. Considering this possibility, there is a risk that the model will seem even to experts to represent Dennett's views without actually doing so.

Finally, questions arise concerning the ability to distinguish humans from replicas. The better such models become, the more difficult it might be to distinguish humans from replicas; outputs of the machine might be mistakenly interpreted as if they were answers of the original human. The results from our planned Turing-test-like experimental investigation will provide first insights into the question of how likely it is that machine outputs are mistakenly taken as answers from the philosopher.[8]

In the distant future, some people might regard the very best replicas as continuations of their personal identity or as having an important part of what matters in identity [23,24].

5. Conclusion

As a preliminary result of our first investigations into how far we can get in creating a digital replica of a philosopher, we can state that our piloting models could produce both surprisingly interesting and totally misguided outputs. Fine-tuning the more powerful Davinci engine did lead to even more impressive outputs but without being able to avoid misguided outputs. Critically acknowledging the misleading output, we conclude that at this stage of research such models are primarily of interest for experts of the work of the author the model is fine-tuned with and for those who are interested in thinking through future possibilities.

Prompt engineering that takes into account the known limitations of such a language model while illustrating the context of a query could significantly increase the proportion of 'better' outputs. However, regardless of how good a model is, it turned out that it is easy to expose it with 'inappropriate' prompts. Nevertheless, the fact that one can expose such models is not necessarily a negative feature because it allows one to clearly distinguish between human and machine outputs. In this respect, it is conceivable that it could be important for potential applications that the user is always made aware that she is interacting with a mindless machine in order to reduce misuse and misunderstanding of how such outputs are to be evaluated from the outset.

So far, our primary comparisons of different strategies regarding fine-tuning indicate that it is important to edit the training data carefully. This included manual editing of the converted pdf templates to fix the errors introduced by the conversion process and deletion of headings, bibliographies, footnotes, images, tables, and the like. Of course, which of these editing decisions had a particular impact on the quality of the output would have to be investigated individually. What seems clear is that redundant sequences, such as recurring headers, page numbers, and lists of references, will probably not contribute to the quality of the model. To this end, we plan to compare our fine-tuned model with models using less edited training data.

We aim to lay the groundwork for the long-term project of creating language-models of prominent philosophers that can be used, transparently and with sensitivity to the difficult ethical issues involved, to gain insight into those philosophers' works by producing outputs that are similar to what those philosophers might have said if prompted with the same questions. Our results so far suggest that this remains a distant dream, and it might require several more radical breakthroughs in technology, but perhaps it will someday be achievable.

References

[1] Rogers A, Kovaleva O, Rumshisky A. A Primer in BERTology: What we know about how BERT works. 2020. Available from: https://arxiv.org/abs/2002.12327
[2] BERT: Official GitHub repository. Available from: https://github.com/google-research/bert
[3] Eleuther AI. Available from: https://www.eleuther.ai
[4] Chowdhery A, Narang S, Devlin J. PaLM: Scaling Language Modeling with Pathways. Google AI Blog. 2022. Available from: https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html
[5] Brown T B, Mann B, Ryder N, Subbiah M, Kaplan, J, Dhariwal P, et al. Language models are few-shot learners. 2020 Apr. Available from: arXiv preprint arXiv:2005.14165
[6] GitHub Copilot. Wikipedia, the free encyclopedia. Available from: https://en.wikipedia.org/wiki/GitHub_Copilot
[7] Langston J. From conversation to code: Microsoft introduces its first product features powered by GPT-3. The AI Blog. 2021. Available from: https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/
[8] Online documentation of Open-AI. Available from: https://beta.openai.com/docs/guides/fine-tuning/case-study-creating-an-expert-model-in-the-legal-domain-which-understands-internal-company-jargon
[9] Online documentation of Open-AI. Available from: https://beta.openai.com/docs/guides/fine-tuning/case-study-maintaining-company-voice
[10] Heaven W D. Open AI's new language generator GPT-3 is shockingly good – and completely mindless. MIT Technological Review. 2020 Jul. Available from: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

[8] For an overview of our preliminary results of this experiment see [22].

[11] Kurenkov A. The inherent limitations of GPT-3. On why GPT-3 as is will not cost many people their jobs or soon lead to AGI. Last week in AI. 2021 Nov. Available from: https://lastweekin.ai/p/the-inherent-limitations-of-gpt-3
[12] Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. Minds & Machines. 2020; 30: 681-94. Available from: https://doi.org/10.1007/s11023-020-09548-1
[13] Bommasani R, Hudson D A, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. 2021 Aug. Available from: https://arxiv.org/abs/2108.07258v1
[14] Christian B. The Alignment problem. Machine Learning and Human Values. New York: W.W. Norton & Company; 2020. 476 p.
[15] Available from: https://dailynous.com/2020/07/30/philosophers-gpt-3/
[16] Schwitzgebel E. More people might soon think robots are conscious and deserve rights. The Splintered Mind. 2021 Mar. Available from: http://schwitzsplinters.blogspot.com/2021/03/more-people-might-soon-think-robots-are.html
[17] Schwitzgebel E. Two robot-generated splintered mind posts. The Splintered Mind. 2021 Nov. Available from: http://schwitzsplinters.blogspot.com/2021/11/two-robot-generated-splintered-mind.html
[18] Online documentation of Open-AI. Available from: https://beta.openai.com/docs/guides/fine-tuning.
[19] Online documentation of Open-AI. Available from: https://beta.openai.com/docs/guides/fine-tuning/preparing-your-dataset
[20] Philosopher-AI. Available from: https://philosopherai.com
[21] Available from: https://www.gov.uk/government/consultations/artificial-intelligence-and-intellectual-property-call-for-views/artificial-intelligence-call-for-views-copyright-and-related-rights
[22] Schwitzgebel, E. Results: The Computerized Philosopher: Can You Distinguish Daniel Dennett from a Computer? The Splintered Mind. 2022 Jul. Available from: http://schwitzsplinters.blogspot.com/2022/07/results-computerized-philosopher-can.html
[23] Schwitzgebel, E. Speaking with the living, speaking with the dead, and maybe not caring which is which. The Splintered Mind. 2021 Jul. Available from: http://schwitzsplinters.blogspot.com/2021/07/speaking-with-living-speaking-with-dead.html
[24] Karpus J, Strasser A. Persons and their digital replicas. (under review).