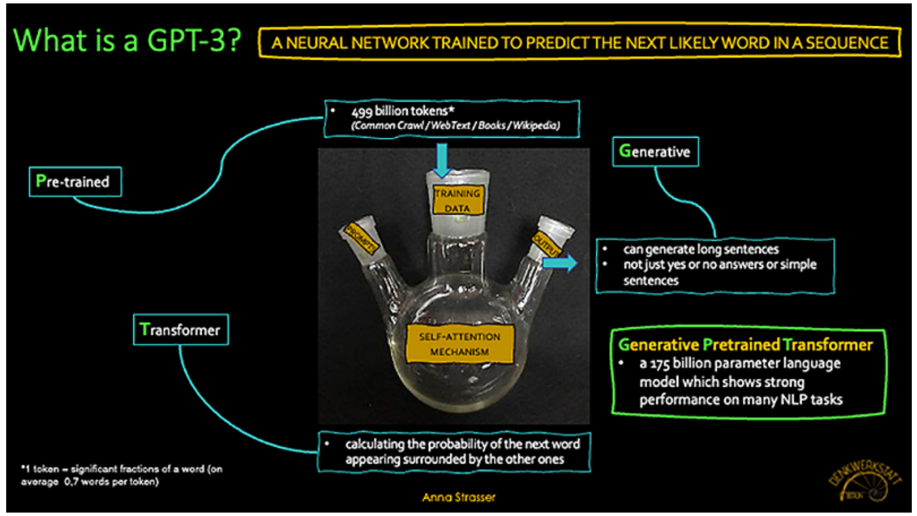
Wie kann man mit ‚intelligenten' Maschinen leben? (Anna Strasser)

Datum: 16.10.2023 (20:00:00–22:00:00)
Ort: Online via Zoom. Für Nicht-Mitglieder von MoMo kann der Link per Email von uns angefragt werden.



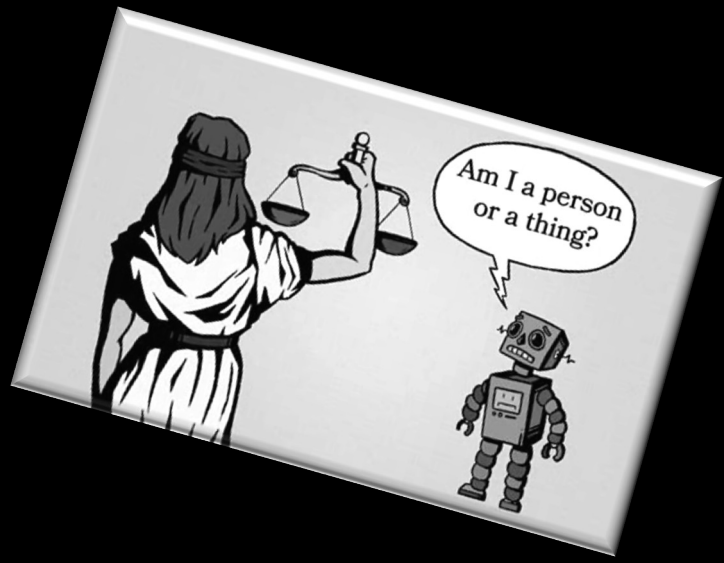
Die Magie ungeheurer statistischer Rechenoperationen

So genannte *Large Language Models* (LLMs) wie ChatGPT haben in letzter Zeit viel Aufmerksamkeit in der Öffentlichkeit bekommen und die KI-Forschung scheint sich auf einen ewigen Sommer zu freuen. Nur was heißt das für uns? Wie werden wir uns in einer Welt zurechtfinden, in der Massen von maschinen-generierten Texten unsere Aufmerksamkeit bekommen und eine Rolle in unserem Leben spielen? Nach dem anfänglichen Enthusiasmus melden sich nun nach und nach kritische Stimmen zu Wort. Ich werde in meinem Vortrag zuerst von einem Projekt berichten, in dem wir ein Sprachmodell namens DigiDan entwickelt haben (Schwitzgebel et al., 2023; Strasser et al., 2023). DigiDan ist ein feingetuntes Sprachmodell, das mit dem Werk von Daniel Dennett trainiert wurde und in der Lage ist, den Anschein zu erwecken, dass es im Namen von Daniel Dennett spricht. Dann werde ich das Publikum einladen, sich mit mir eine Welt vorzustellen, in der es zu jedem Autor ein Chatbot gibt, die Internetseiten voll von maschinen-generierten Texten sind und viele wichtige Entscheidung KI-unterstützt vollzogen werden. Eine Analyse von diesem Szenario wird zeigen, wie destruktiv diese neue Technologie sein kann und wie schwierig es ist, sie angemessen zu regulieren (Strasser, 2023).

Bibliographische Referenzen:

- Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). *Creating a large language model of a philosopher.* Mind & Language, n/a(n/a). https://doi.org/10.1111/mila.12466
- Strasser, A. (2006). *Kognition künstlicher Systeme. In Kognition künstlicher Systeme*. De Gruyter. https://doi.org/10.1515/9783110321104
- Strasser, A. (2023). *On pitfalls (and advantages) of sophisticated large language models* (arXiv:2303.17511). arXiv. https://doi.org/10.48550/arXiv.2303.17511
- Strasser, A., Crosby, M., & Schwitzgebel, E. (2023). *How Far Can We Get in Creating a Digital Replica of a Philosopher?* In: R. Hakli, P. Mäkelä, & J. Seibt (Eds.), Social Robots in Social Institutions (pp. 371–380). IOS Press. https://doi.org/10.3233/FAIA220637
- Strasser, A., Sohst, W., Stepec, K., & Stapelfeldt, R. (Eds.). (2021). *Künstliche Intelligenz – Die große Verheißung.* (Reihe MoMo KonText, Vol. 8). xenomoi Verlag.

Anna Strasser about LLMs

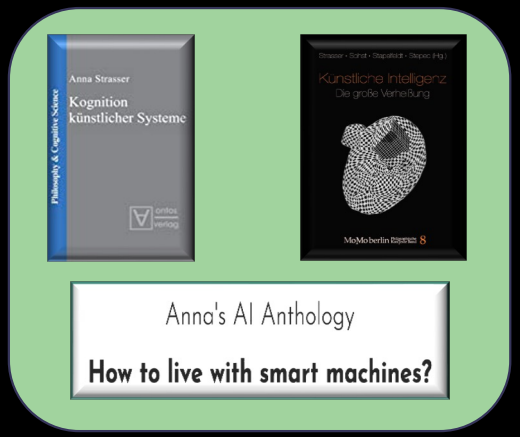


DENKWERKSTATT BERLIN

So genannte *Large Language Models* (LLMs) wie ChatGPT haben in letzter Zeit viel Aufmerksamkeit in der Öffentlichkeit bekommen und die KI-Forschung scheint sich auf einen ewigen Sommer zu freuen.

NUR WAS HEISST DAS FÜR UNS?

Wie werden wir uns in einer Welt zurechtfinden, in der Massen von maschinen-generierten Texten unsere Aufmerksamkeit bekommen und eine Rolle in unserem Leben spielen?



- Daniel Dennett: We are all Cherry-Pickers
- Eric Schwitzgebel & Anna Strasser: Asymmetric joint actions
- David Chalmers: Do large language models extend the mind?
- Henry Shevlin: LLMs, Social AI, and folk attributions of consciousness
- Keith Frankish: What are large language models doing?
- Joshua Rust: Minimal Institutional Agency
- Ned Block: Large Language Models are more like perceivers than thinkers
- Paula Droege: Full of sound and fury, signifying nothing
- https://youtube.com/playlist?list=PL-ytDJty9ymIBGQ7z5iTZjNqbfXjFXI0Q&si=noZ7bPGz-uMt6jmm

DigiDan

Szenario

Analyse

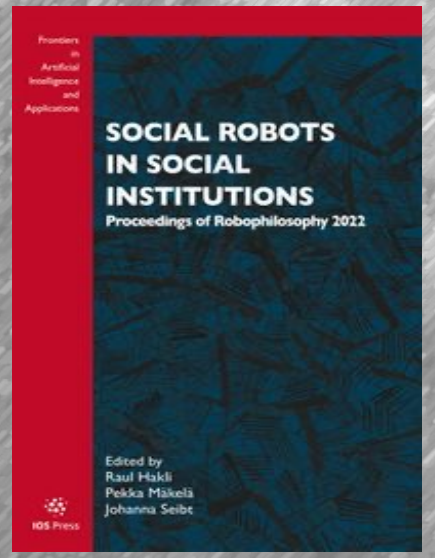
Strasser, A., Crosby, M., & Schwitzgebel, E. (2023).
How Far Can We Get in Creating a Digital Replica of a Philosopher?
In R. Hakli, P. Mäkelä, & J. Seibt (Eds.), *Social Robots in Social Institutions* (pp. 371–380). IOS Press. https://doi.org/10.3233/FAIA220637

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023).
Creating a large language model of a philosopher.
Mind & Language. https://doi.org/10.1111/mila.12466

einer zukünftigen Welt, in der

- es zu jedem Autor ein Chatbot gibt
- die Internetseiten voll von maschinen-generierten Texten sind
- viele wichtige Entscheidung KI-unterstützt vollzogen werden

- wie disruptive diese neue Technologie sein kann
- wie schwierig es ist, sie angemessen zu regulieren (Strasser, 2023)



(Strasser et al., 2023)

THE MAKING OF DIGIDAN

LLM made a first impressive appearance

LARGE LANGUAGE MODELS (LLMs)

generating long strings of text in response to a prompt

NEURAL NETWORKS | UNSUPERVISED MACHINE LEARNING | SELF-ATTENTION MECHANISM → TRANSFORMERS

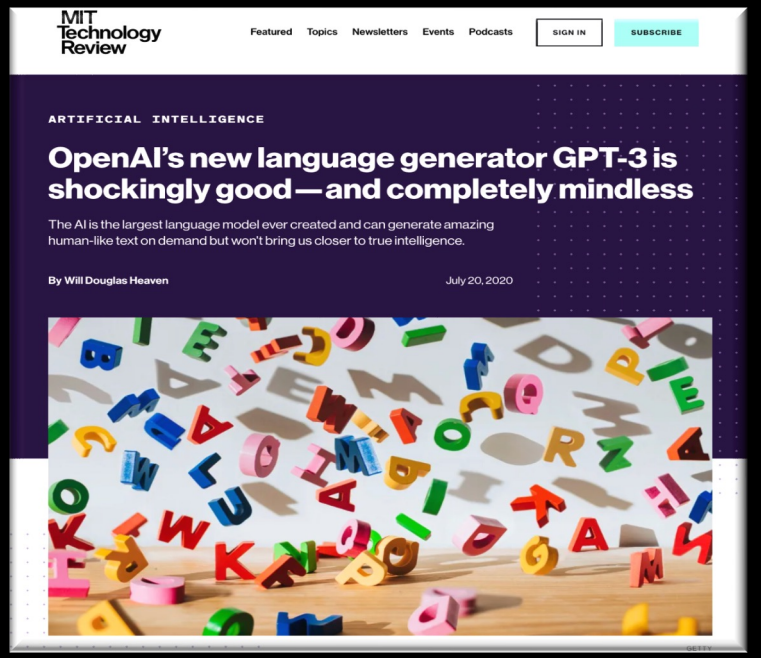
With such machines, you can engage in seemingly intelligent conversations.

- e.g., if you ask a question, the machine will often (not always) generate a sensible-seeming answer.

notable successes in many domains

- 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 (LLMs)

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



(Heaven, 2020)

What is a GPT-3?

A NEURAL NETWORK TRAINED TO PREDICT THE NEXT LIKELY WORD IN A SEQUENCE

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

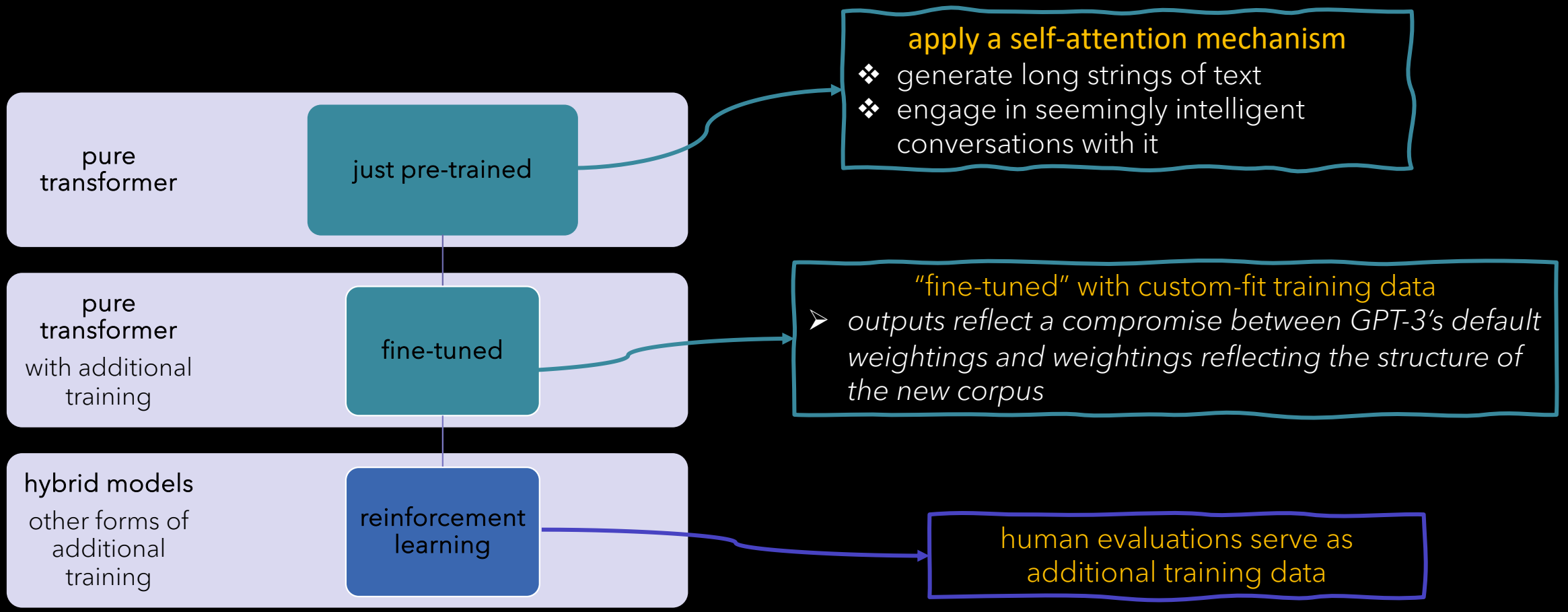
- a 175 billion parameter language model which shows strong performance on many NLP tasks

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

Anna Strasser

DENKWERKSTATT BERLIN

OTHER LARGE LANGUAGE MODELS



AI 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?

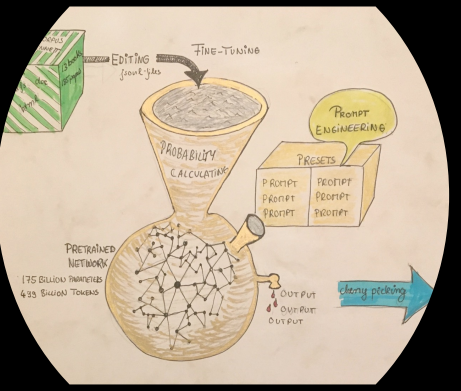
Daniel Dennett

Eric Schwitzgebel

Mathew Crosby

WITH DANIEL DENNETT'S PERMISSION, WE FINE-TUNED AN LLM WITH THE CORPUS 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.

DigiDan

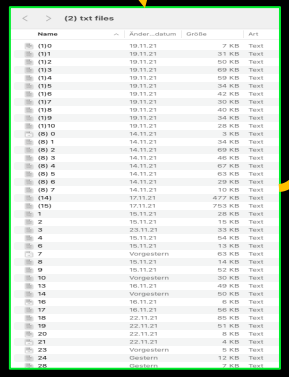


Strasser, Schwitzgebel & Crosby 2023; Schwitzgebel et al. 2023

Editing & fine-tuning

Preparing training data

Dennett's corpus



converted into plain text format
- stripping away headers, footnotes, scanning errors, marginalia, and other distractions

jasonl training data

Dinner is ready! Today we serve three million tokens

15 BOOKS
269 ARTICLES

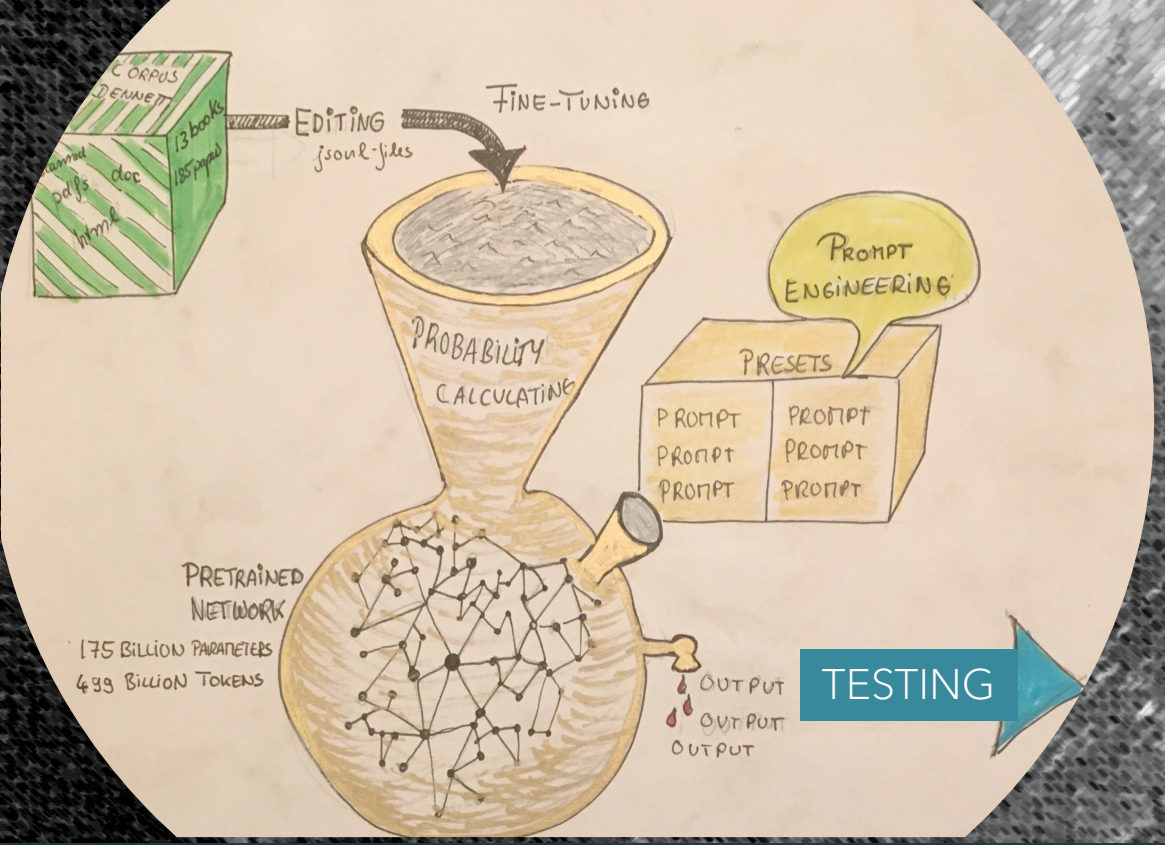
Segments of training data (<2000 tokens)

Blank prompts

```
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



(Schwitzgebel et al., 2023)

TESTING DIGIDAN

Testing DigiDan

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

significantly below the hypothesized accuracy of 80%

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

Guessing task & Evaluation of the likeliness

1

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:

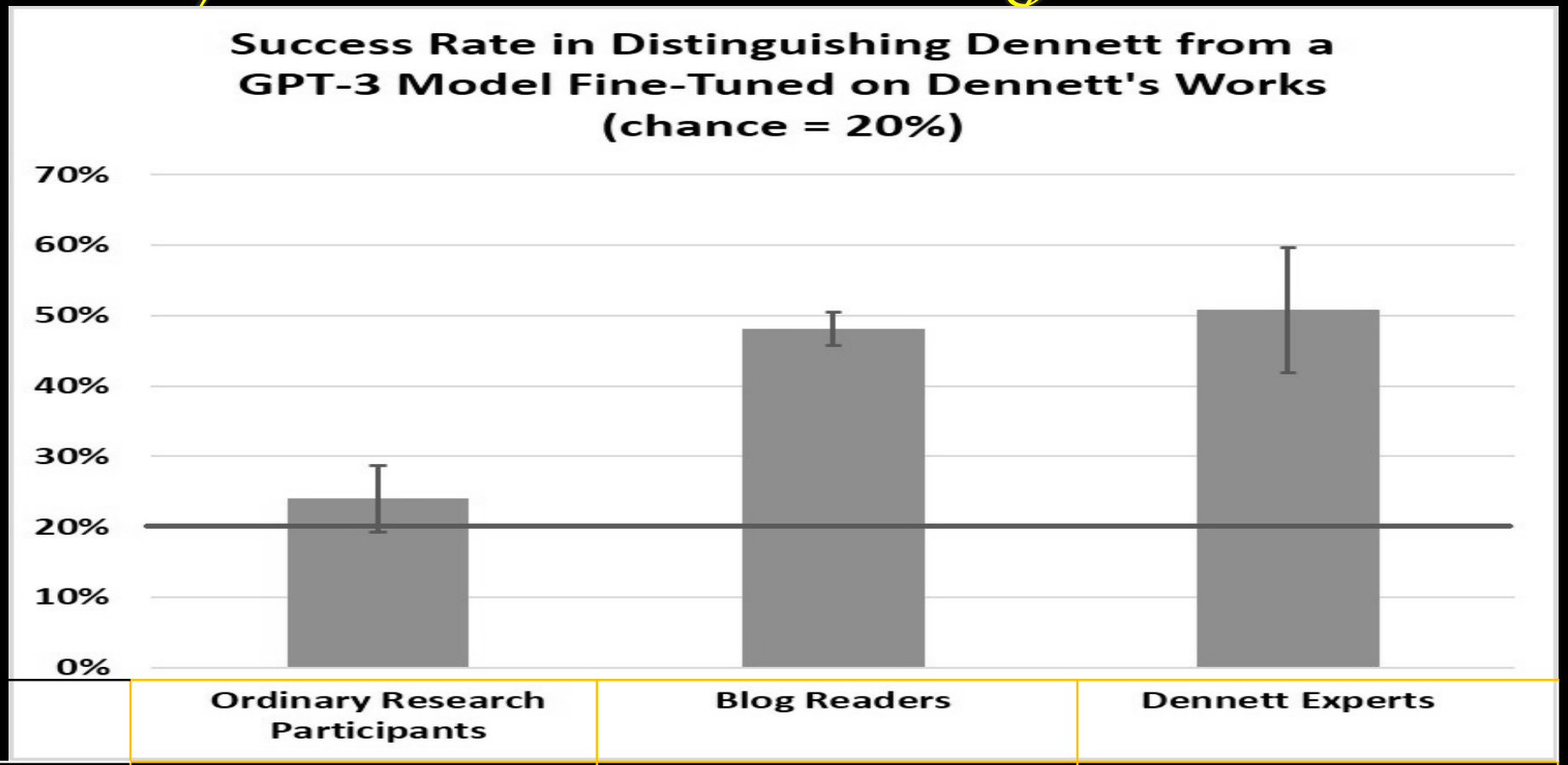
2

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)

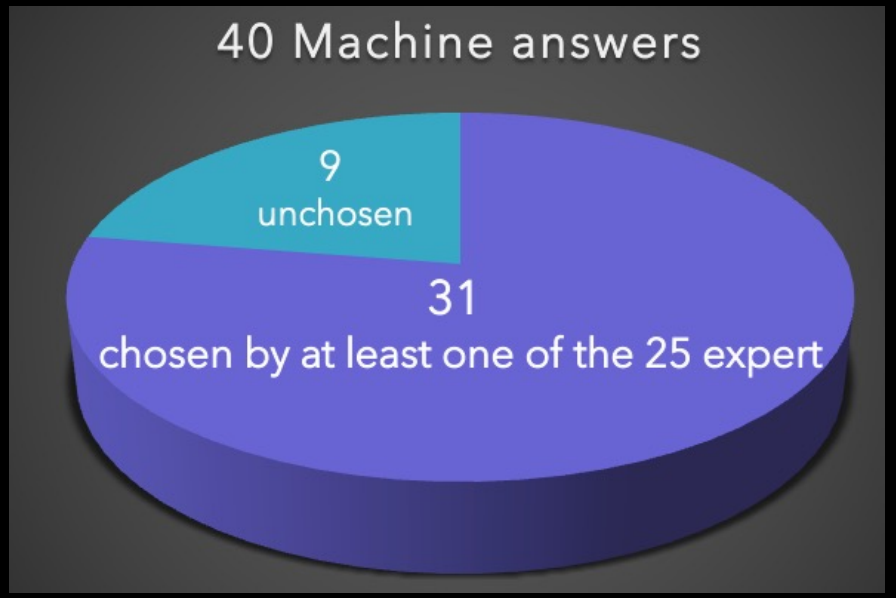
DENKWERKSTATT BERLIN

DigiDan was much better than expected



majority	with no classes in philosophy & no familiarity with Dennett's work	with graduate degrees in philosophy & familiarity with Dennett's work	reported having read over 1000 pages of Dennett's work
correctly guessed	1.20 times out of 5 • 86% 1-2 correct • 14% 3-4 correct	4.81 times out of 10 (48%)	5.08 times out of 10 (51%)
given a five-alternative forced choice	• near chance rate of 20%	• substantially above chance	

but not reliable!!!



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

I asked myself whether I was happy that I got involved in this project.

‘TALKING’ TO PHILOSOPHERS IS QUITE ATTRACTIVE THAN ONLY INTERPRETING THEIR TEXTUAL OUTPUTS

BLURRING THE DIFFERENCE BETWEEN HUMANS & MACHINES

I hope no language model will ever be trained with all my statements.

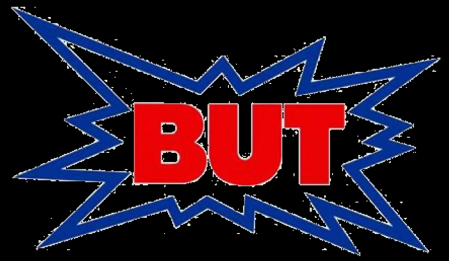
- I do not aim to be mistaken for such a model.
- I do not aim to have such a digital legacy continuing to make statements on my behalf after my death.

DENKWERKSTATT BERLIN

Hard to distinguish

Just ten years ago

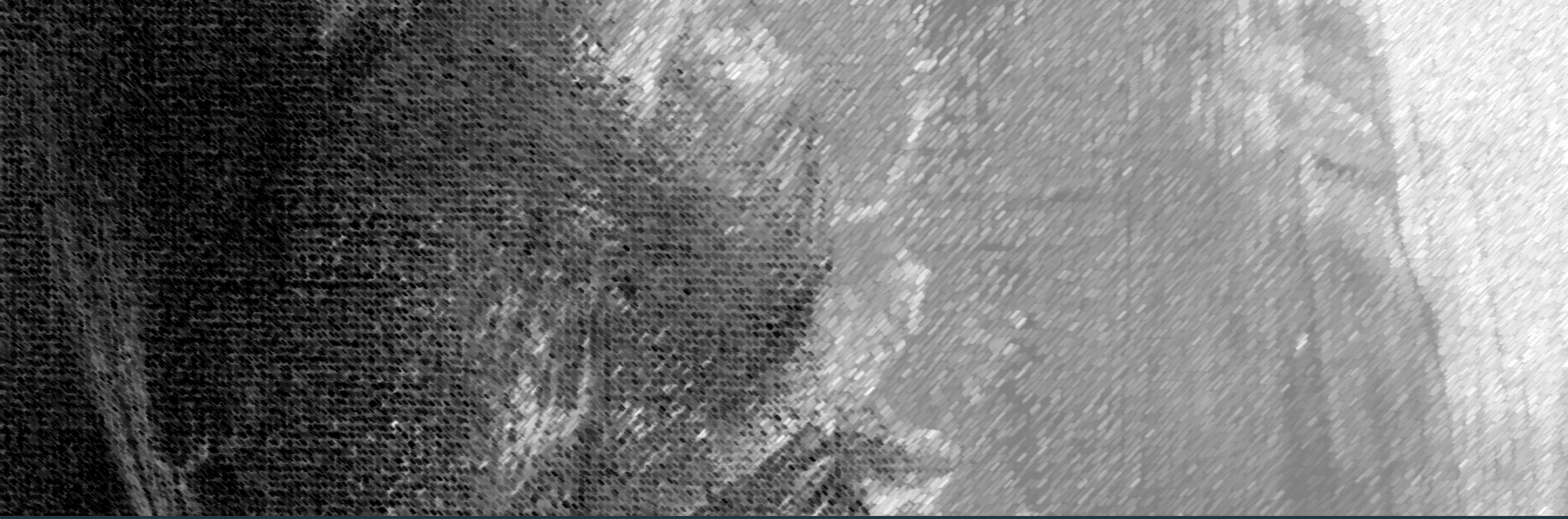
- nobody worried about their abilities to distinguish between human-made & machine-generated text
- differences were so obvious
  - it didn't seem like that would change quickly



THIS HAS CHANGED RIGIDLY

We all should be worried because neither humans nor sophisticated detection software can distinguish with certainty between human-generated and machine-generated text

THE INCREASING INDISTINGUISHABILITY HAS THE POTENTIAL TO CONTRIBUTE TO AN EPISTEMOLOGICAL CRISIS.



IMAGINE

eine Welt, in der

- es zu jedem Autor ein Chatbot gibt
- die Internetseiten voll von maschinen-generierten Texten sind
- viele wichtige Entscheidung KI-unterstützt vollzogen werden

How can we trust the content of websites?

How you decide whether you trust the content of websites?

VISITING A WEBSITE FROM *STANFORD ENCYCLOPEDIA*

you

- trust that all those articles are written by scientific scholars
- rely on their expertise
- belief that cited references are existing
- assume that the articles went through a reviewing process

FINE-TUNED LLM THAT CAN PRODUCE HARD TO DISTINGUISHABLE CONTENT

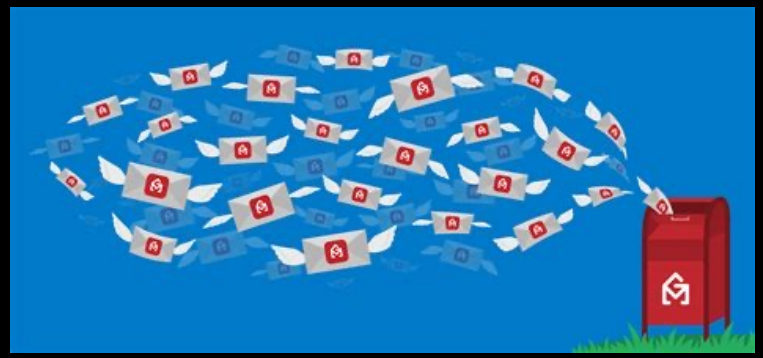
- article may contain a number of serious flaws
  - hallucinated references
  - paraphrases concerning position of other philosophers that are just wrong
- you would have to doublecheck everything

- And maybe there is another LLM that is compiling all the papers of the hallucinated references …
- no chance to find out whether you can trust that information

… unless you go back to a library and check in real books and journals

Responsibility in Human-Machine Interaction

WORKING WITH A MODERN AI SYSTEM ON A MORALLY DELICATE TASK



AI system trained by a human that is able to compose and send 100s of emails at a go using prepared information about the recipients

JOINT TASK

sending emails to inform vulnerable people of either good or bad news that will change their lives unalterably with sharp time constraints

AI:

- composing & sending emails

HUMAN:

- monitor the emails
  - at a rate that is acceptable for the job
  - but also distracted by other things

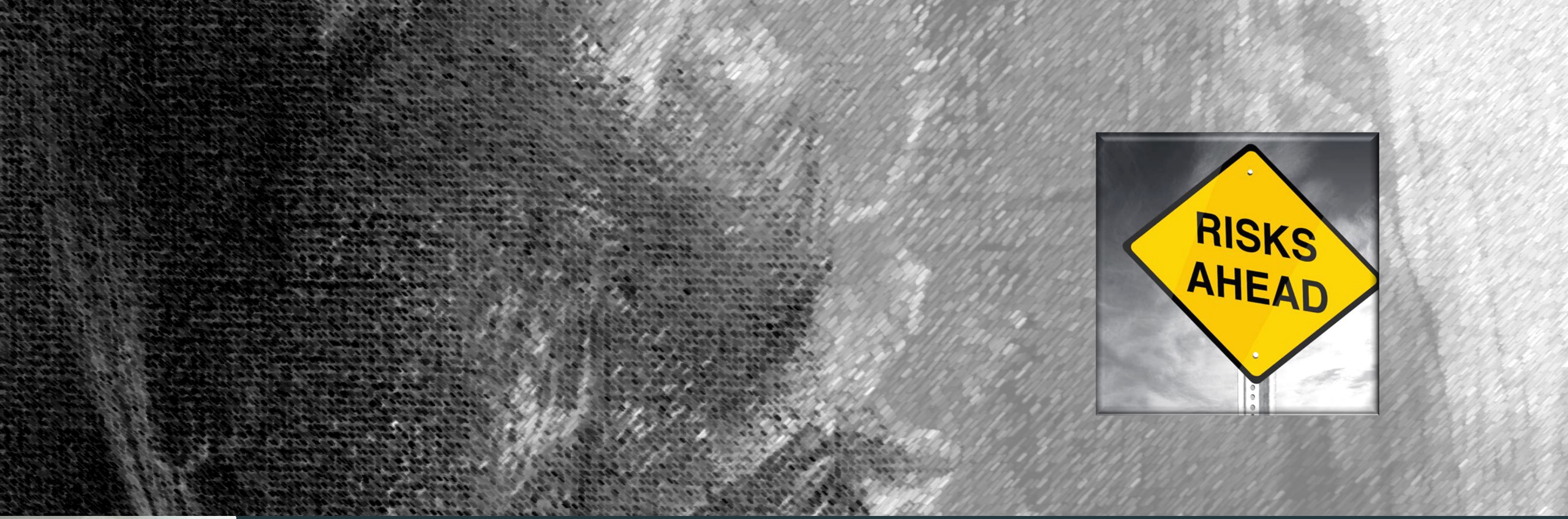
SOMETHING GOES WRONG

- several of the batches of emails from the AI sending out inappropriate and inaccurate information

→ **going to be highly harmful & hurtful to the individuals**

➢ Significant harms have been incurred.

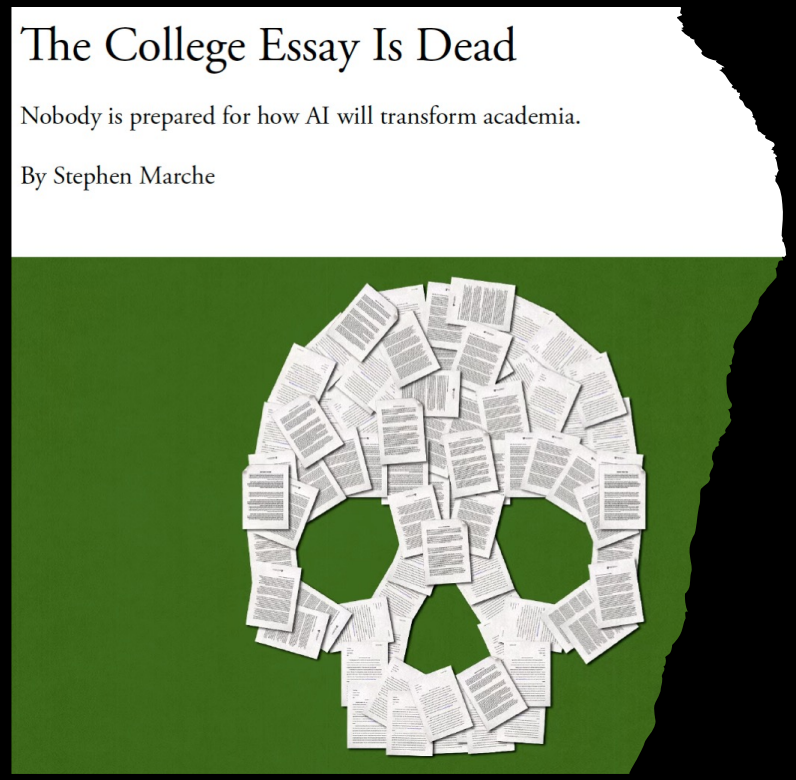
DENKWERKSTATT BERLIN



ANALYSIS

Authorship

How to deal with verifiable authorship with respect to the mass of electronically distributed texts?



- students might soon have a hard time proving their authorship when sending in their essays
- teachers might not be sure whether they are not grading the outputs of an LLM

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

DENKWERKSTATT BERLIN

Strasser, Anna (2023). On pitfalls (and advantages) of sophisticated Large Language Models. preprint at https://arxiv.org/abs/2303.17511

Copyright

Copyright law governing the use of training-data is not yet settled

- unclear whether it is fair use of intellectual property
- open question
  - How to deal with works by deceased authors? (Nakagawa & Orita 2022)

PRESS RELEASES

More than 15,000 Authors Sign Authors Guild Letter Calling on AI Industry Leaders to Protect Writers

Artificial Intelligence

July 18, 2023

https://authorsguild.org/news/thousands-sign-authors-guild-letter-calling-on-ai-industry-leaders-to-protect-writers/

https://www.deutschlandfunkkultur.de/hollywood-schauspieler-streik-100.html

How film studios plan to use AI systems in the future

- actors fear being replaced by AI-generated avatars created by scans
- writers fear studios will soon have scripts written by AI software.

➢ There are calls for clear rules for the use of artificial intelligence in film and series production.

DENKWERKSTATT BERLIN

ChatGPT's hallucination just got OpenAI sued. Here's what happened

https://www.zdnet.com/article/chatgpts-hallucination-just-got-openai-sued-heres-what-happened/

Overreliance

Fine-tuned language models can create opportunities for over-reliance

- Our efforts to make sense of anything that looks roughly interpretable can betray us!

NOT GOOD ENOUGH!

DigiDan did not reliably produce outputs representing Dennett's views.

not surprising:

- all deep learning networks 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

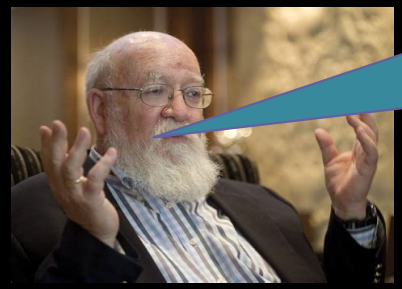
I RECOMMEND SUBSTANTIAL CAUTION
BEFORE RELEASING TO THE PUBLIC ANY LANGUAGE MODELS FINE-TUNED ON AN INDIVIDUAL AUTHOR

Counterfeits

Creating counterfeit digital people risks destroying our civilization. Democracy depends on the informed (not misinformed) consent of the governed. By allowing the most economically and politically powerful people, corporations, and governments to control our attention, these systems will control us. Counterfeit people, by distracting and confusing us and by exploiting our most irresistible fears and anxieties, will lead us into temptation and, from there, into acquiescing to our own subjugation. the counterfeit people will talk us into adopting policies and convictions that will make us vulnerable to still more manipulation. Or we will simply turn off our attention and become passive and ignorant pawns. This is a terrifying prospect. (Dennett 2023)

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



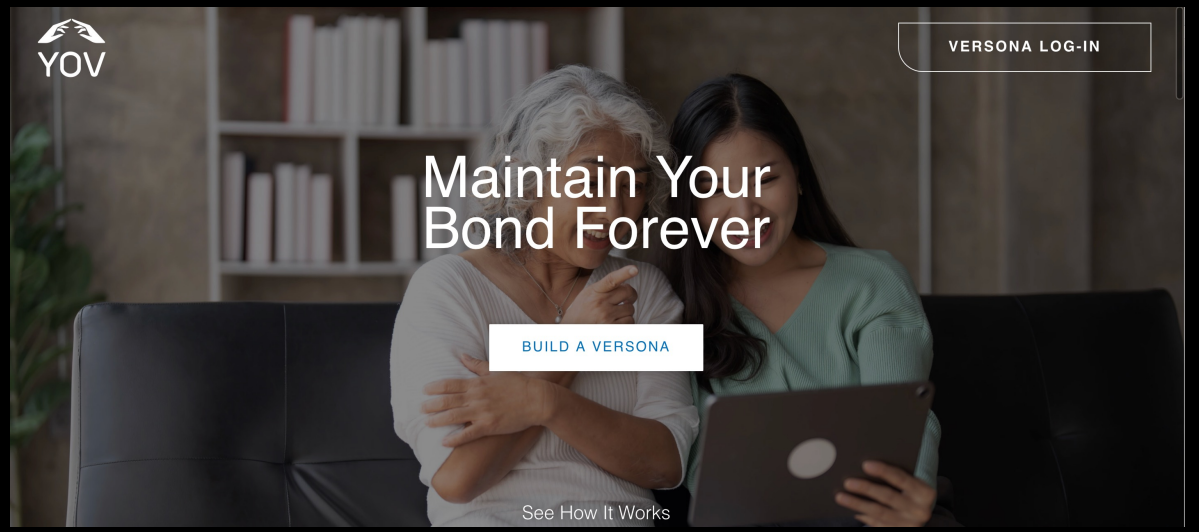
Dennett as interviewed in Cukier 2022

COUNTERFEITING IS A SERIOUS ACT OF SOCIAL VANDALISM

https://youtu.be/GzSFn4FCGgI?si=acDDNieRmROmpi42

DENKWERKSTATT BERLIN

Digital replicas



https://www.myyov.com

Aufgrund aller möglichen Fälschungen ist eine erkenntnistheoretische Krise zu erwarten, und die Menschen werden darauf achten müssen, was sie für eine echte Person halten.

Um zu vermeiden, dass wir zu misstrauisch und paranoid werden, brauchen wir Gesetze die regeln, wie sich KI-produzierte Outputs präsentieren, und wir werden wahrscheinlich neue Strategien entwickeln müssen, um unsere Gegenüber als Menschen zu erkennen.

DENKWERKSTATT BERLIN

Take home message

EMPFEHLUNGEN

Wir brauchen eine Gesetzgebung, die einige der Verwendungsmöglichkeiten dieser Systeme verbietet!

Wir sollten immer um Erlaubnis bitten, wenn wir ein Modell auf der Grundlage einer lebenden Person bauen!

DENKWERKSTATT BERLIN

Jetzt seid Ihr an der Reihe,
darüber nachzudenken,
wie wir mit der zunehmenden
Ununterscheidbarkeit
umgehen könnten.

All this would not have been possible if I had not interacted with people & machines

Daniel Dennett

Eric Schwitzgebel

Mathew Crosby

David Schwitzgebel

Jurgis Karpus

Mike Wilby

- Schwitzgebel, Eric, Schwitzgebel, David, Strasser, Anna (2023). **Creating a Large Language Model of a Philosopher.** Mind & Language, 1–22. https://doi.org/10.1111/mila.12466

open access

- Strasser, A., Crosby, M., Schwitzgebel, E. (2023). **How far can we get in creating a digital replica of a philosopher?** In R. Hakli, P. Mäkelä, J. Seibt (eds.), *Social Robots in Social Institutions. Proceedings of Robophilosophy 2022.* Series Frontiers of AI and Its Applications, vol. 366, 371–380. IOS Press, Amsterdam. doi:10.3233/FAIA220637

- Strasser, A., Wilby, M. (2023). **The AI-Stance: Crossing the Terra Incognita of Human-Machine Interactions?** In R. Hakli, P. Mäkelä, J. Seibt (eds.), *Social Robots in Social Institutions. Proceedings of Robophilosophy 2022.* Series Frontiers of AI and Its Applications, vol. 366, 286–295. IOS Press, Amsterdam. doi:10.3233/FAIA220628

Thank you !

DigiDan