ESPP 2023

What shall we do

with the increasing indistinguishability between machines and humans?

Anna Strasser

DenkWerkstatt Berlin / LMU Munich

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

BUT

THIS HAS CHANGED RIGIDLY

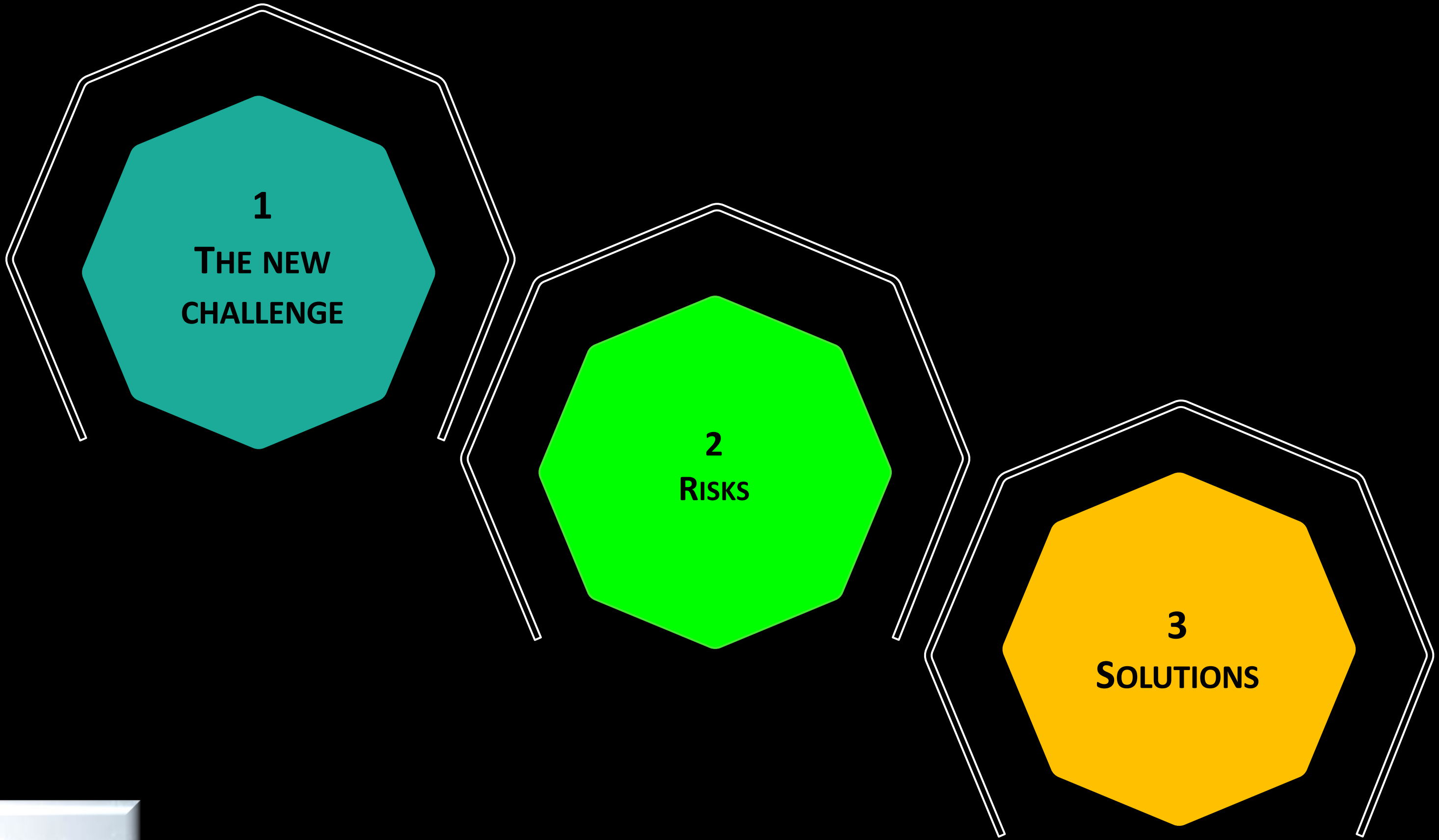
We all should be worried because neither humans nor sophisticated detection software can distinguish with certainty between human-generated and machine-generated text

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

DENKWERKSTATT
BERLIN

Overview

slides can be downloaded at
https://www.denkwerkstatt.berlin/ANNA-STRASSER/TALKS/



1
The new challenge

2
Risks

3
Solutions

1

The new challenge

LLM made a first impressive appearance

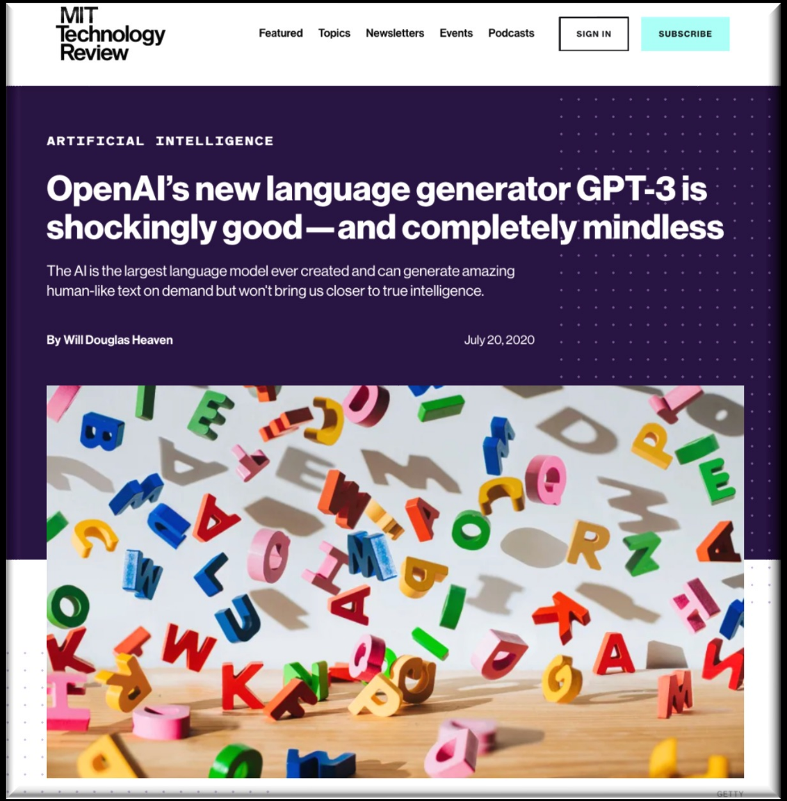
LARGE LANGUAGE MODELS (LLMs)

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

- *generating long strings of text in response to a prompt*

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.



ARTIFICIAL INTELLIGENCE

OpenAI's new language generator GPT-3 is shockingly good—and completely mindless

The AI is the largest language model ever created and can generate amazing human-like text on demand but won't bring us closer to true intelligence.

By Will Douglas Heaven

July 20, 2020

(Heaven, 2020)

DENKWERKSTATT BERLIN

GPT-3 is a large language model

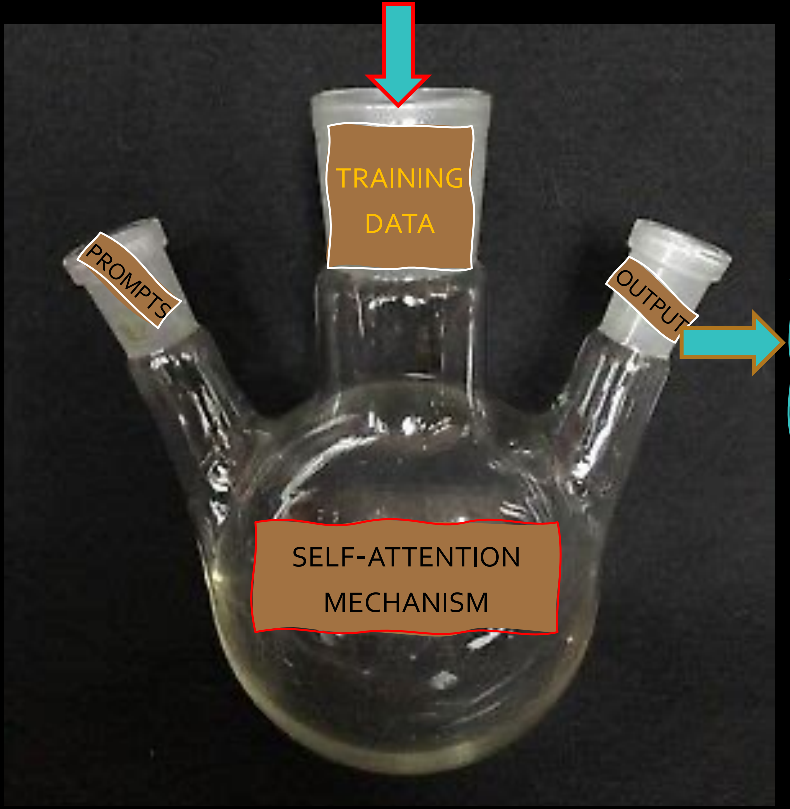
a neural network trained to predict the next likely word

Pre-trained

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

Generative

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



Wolfram, S. (2023). What Is ChatGPT Doing ... and Why Does It Work.

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

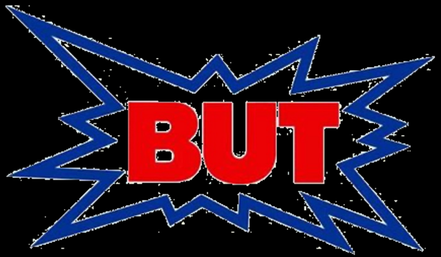
*1 token = significant fractions of a word (on average 0,7 words per token)

AI research has made huge progress

NOTABLE SUCCESSES IN MANY DOMAINS

- producing original prose with fluency equivalent to that of a human (LLMs)
- discovering novel algorithms, protein folding (AlphaTensor, AlphaFold)
- automatic translation (DeepL)
- computer code generation (Github Copilot)
- …

Jumper, Evans, & Pritzel et al. 2021; Fawzi et al. 2022; Steven & Iziev 2022



I take a critical stance, especially towards the quality of LLMs performance that we can observe in conversation-like situations and in situations in which they are used to gain knowledge.

AI can outperform even expert humans in many domains

successes in discovering novel algorithms, protein folding, automatic translation, computer code generation, and producing original prose with fluency equivalent to that of a human

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?

DigiDan

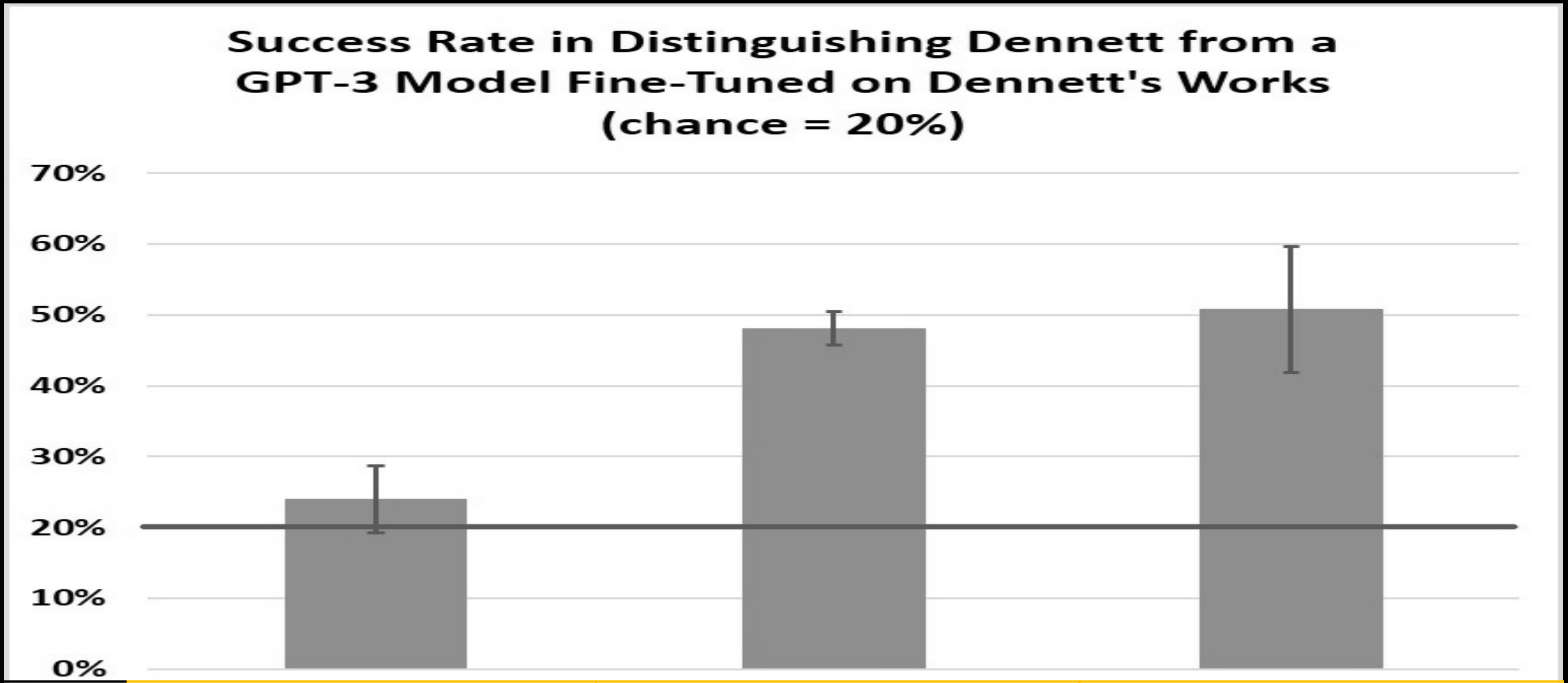
With Daniel Dennett's permission, we fine-tuned an LLM with the corpus of Daniel Dennett sufficiently good that experts in Dennett's work could not reliably distinguish paragraphs written by Dennett from those written by the language model.

Our experiment testing the discrimination abilities might be taken as an indirect measure of the quality of the performance of our model.

DENKWERKSTATT BERLIN

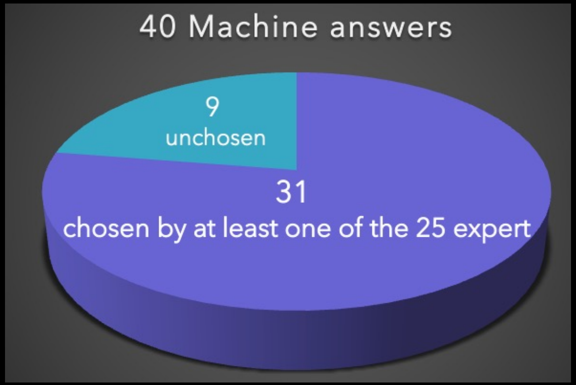
Strasser, Schwitzgebel & Crosby 2022; Schwitzgebel et al. 2023

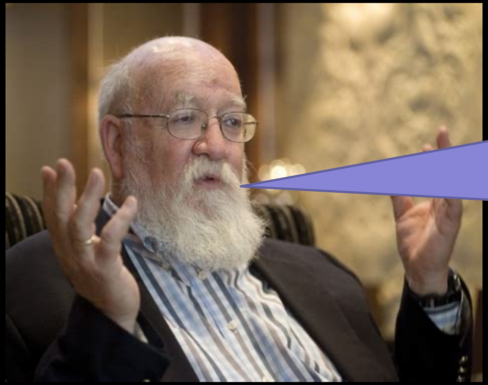
DigiDan was much better than expected



	Ordinary Research Participants	Blog Readers	Dennett Experts
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	

Performance of the machine answers





> *"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

overall performance is not reliable → do not over-rely on such models

ALL LLMS WHICH ARE BASED ON NEURAL NETWORKS COME WITH LIMITATIONS REGARDING RELIABILITY.

- produce unhuman-like mistakes
- inconsistent in their outputs
- hallucinate facts

Human discrimination abilities

informal assessments showing that it is hard to distinguish
(Rajnerowicz 2022; Sinapayen 2023; Vota 2020)

other studies using psychological methods to test humans' discrimination abilities

e.g., Clark et al. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text

difference between GPT-2 & GPT-3

- texts in 3 domains: stories, news articles, recipes
- 5 selected texts → judge whether these texts were likely to have been generated by humans or by machines

Results	GPT-2:	GPT-3:
accuracy in discriminating	58% significantly above chance	only 50% not significantly different from chance

➢ scaling up the models makes it more difficult to distinguish

THE MORE ADVANCED LLMS ARE, THE MORE DIFFICULT IT BECOMES TO DISTINGUISH BETWEEN MACHINE-GENERATED & HUMAN-MADE TEXT.

Discrimination with the help of detection software

BUT detection software cannot distinguish with 100% certainty between machine-generated & human-made text

two types of errors:

1. false-negative (machine-generated text falsely judged to be written by humans)
2. false-positive (human-generated text falsely judged to be machine-generated)

ARMS RACE BETWEEN FRAUDSTERS & FRAUD DETECTION

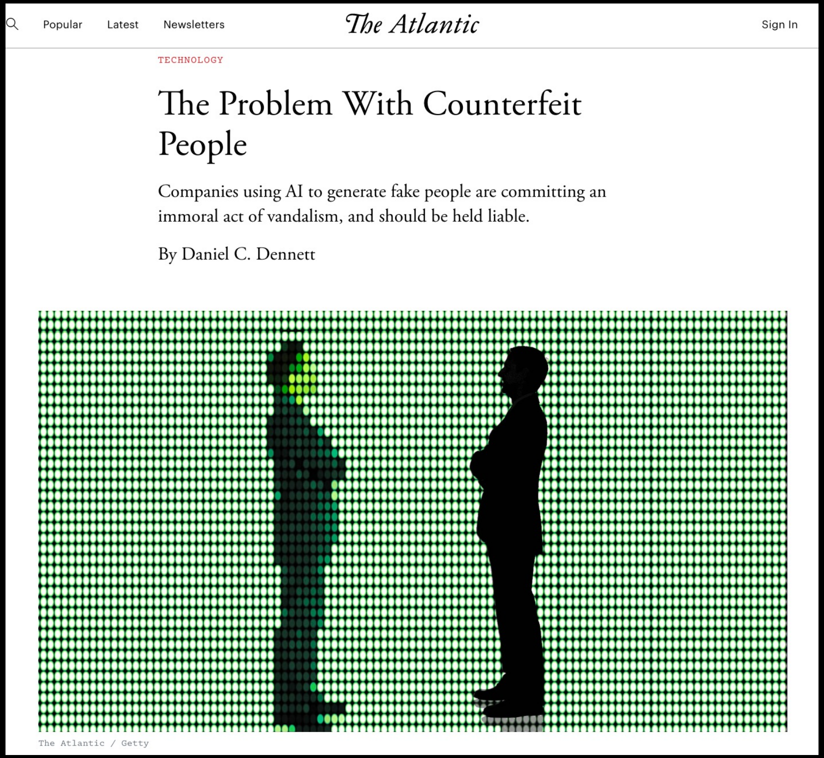
2

Risks

DENKWERKSTATT

BERLIN

Counterfeits



Popular Latest Newsletters *The Atlantic* Sign In

TECHNOLOGY

The Problem With Counterfeit People

Companies using AI to generate fake people are committing an immoral act of vandalism, and should be held liable.

By Daniel C. Dennett

The Atlantic / Getty

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)

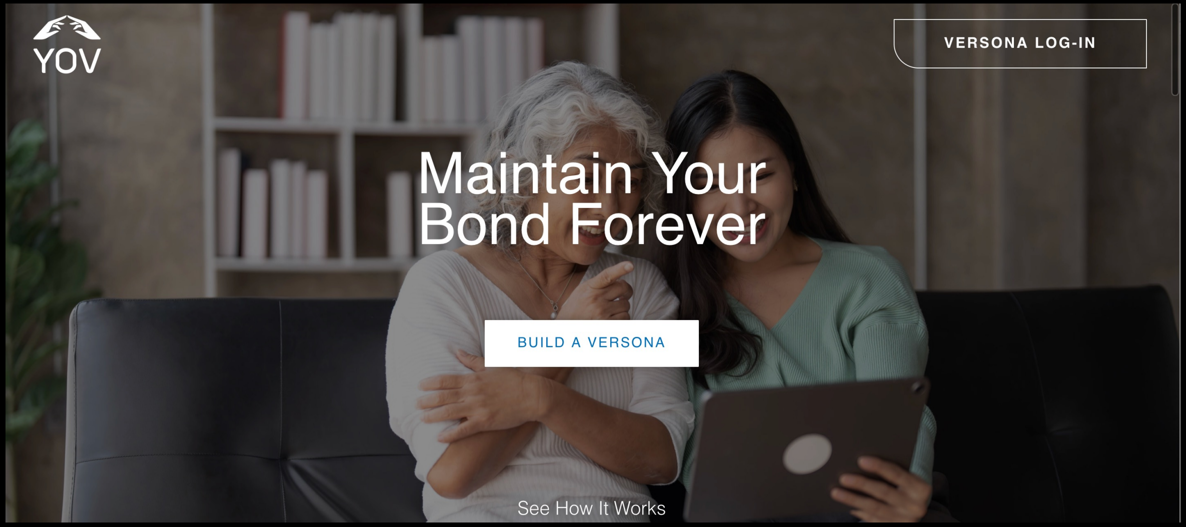
https://youtu.be/GzSFn4FCGgI?si=acDDNieRmROmpi42

COUNTERFEITING IS A SERIOUS ACT OF SOCIAL VANDALISM

Digital replicas

Karpus, Jurgis & Strasser, Anna (submitted). Persons and their digital replicas

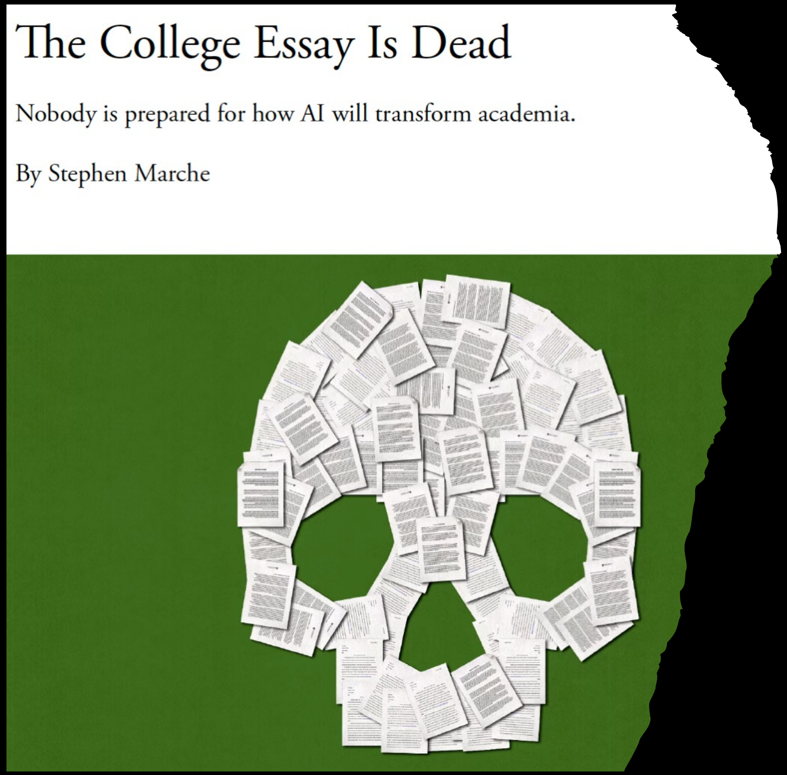
‘Be right back’ of the Black Mirror TV series



https://www.myyov.com

Authorship

HOW TO DEAL WITH VERIFIABLE AUTHORSHIP WITH RESPECT TO THE MASS OF ELECTRONICALLY DISTRIBUTED TEXTS?



The College Essay Is Dead

Nobody is prepared for how AI will transform academia.

By Stephen Marche

- 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

- researchers might soon have a hard time proving their authorship
- publishing houses may not be able to avoid publishing machine-generated papers

- 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

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

DENKWERKSTATT BERLIN

Due to all potential deep fakes, there is an epistemological crisis to be expected, and people will need to look out for what they take as representing a real person.

Avoiding that we get too suspicious and paranoid, we might need new laws for how AIs present themselves, and we will probably have to develop new strategies for identifying our counterparts as humans.

3

SOLUTIONS

DENKWERKSTATT BERLIN

Regulation & punishment

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

DENKWERKSTATT BERLIN

The new EU AI Act

MOST BASIC REQUIREMENT

AI IS ALWAYS IDENTIFIED → NO ONE THINKS THEY ARE TALKING TO A PERSON WHEN THEY REALLY ARE TALKING TO A MACHINE

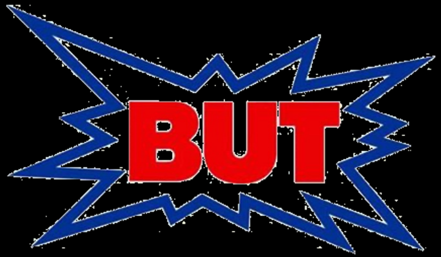
But how can we check whether people follow this law if we cannot distinguish human-made from machine-generated text?

DIGITAL WATERMARKS

(Wigger 2022)

Kirchenbauer et al. (2023)

- require the creators of LLMs to add a watermark signal to each generated text passage
  - that cannot be easily removed by simply modifying the text
- provide open-source software for watermark detection



- not all LLM creators will adhere to it
- possible to fool watermark detectors.

ARMS RACE BETWEEN FRAUDSTERS AND THOSE WHO WANT TO MARK LLM’S OUTPUTS RECOGNIZABLY.

Is there a solution?

IFF THERE IS NO COMPLETELY RELIABLE METHOD FOR DETECTING AI-GENERATED TEXT?
WHAT SHOULD WE DO?

Bans cannot be enforced proactively, which means that one has to rely on human help.

IT SEEMS AS IF WE ARE NOT PREPARED FOR THE EMERGENCE OF SUCH DISRUPTIVE AND NOVEL TECHNOLOGIES.
WHAT CAN WE HOPE FOR?

HOPE 1: humans make mistakes as well

😇 limitations concerning the reliability might not be that awful in the future?
😇 maybe we will have reasons to trust future machine-generated text more than we can do right now?

HOPE 2

😇 we live already with a lot of technology that can be misused

Will we establish new social practices that aim at proving that one is really the original author of what is written or said?

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

- Perhaps universities will return to supervised essay writing in person.
- Any time a detection algorithm or a teacher accuses a text to be machine-generated the author is invited to a face-to-face conversation to defend their authorship

A new social practice

A more minimal notion of moral blame

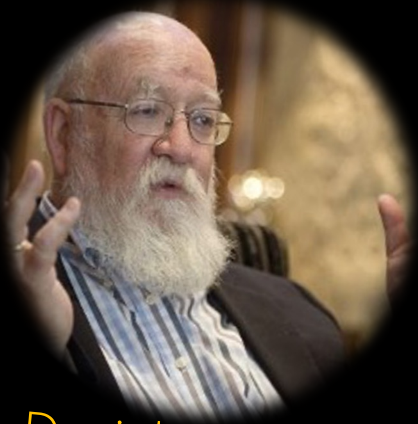
- involves just the behavioral component, which Scanlon calls a “modification” of the relationship, involving the “withdrawal of trust” (Scanlon 2015, p. 93).

Applying this to AI systems

- we suggest that a new social practice
  → a normatively appropriate ‘withdrawal of trust’ presupposing that after each human-machine interaction, there will be a procedure of evaluation and the human is responsible for checking whether the AI did learn from this evaluation.
- And this can serve as a basis of withdrawal of trust.

Now, it is your turn to think about how we can handle the increasing indistinguishability.

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



Daniel Dennett

Eric Schwitzgebel

Mathew Crosby

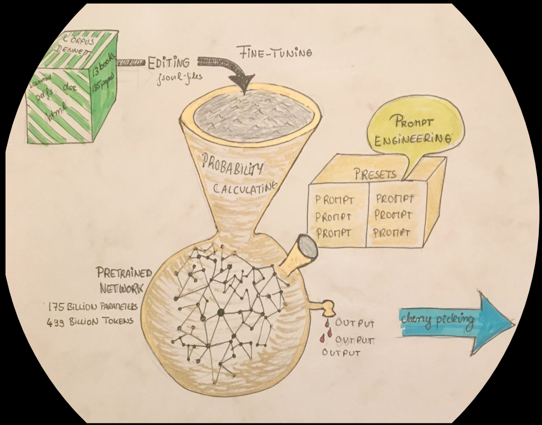
Thank you!

David Schwitzgebel

Jurgis Karpus

Mike Wilby

DigiDan



REFERENCES

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). *All That's "Human" Is Not Gold: Evaluating Human Evaluation of Generated Text* (arXiv:2107.00061). arXiv. https://doi.org/10.48550/arXiv.2107.00061

Dennett, D. C. (2023). *The Problem With Counterfeit People*. The Atlantic. https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/

Dennett, D. C. (2023). *We are all cherry-pickers.* https://youtu.be/GzSFn4FCGgI?si=acDDNieRmROmpi42

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

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

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

Heaven, W. D. (2020). *OpenAI's new language generator GPT-3 is shockingly good—And completely mindless*. MIT Technology Review. https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

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

Karpus, J. & Strasser, A. (submitted). Persons and their digital replicas.

Rajnerowicz, K. (2022).·Human vs. AI Test: Can We Tell the Difference Anymore? Statistics & Tech Data Library. https://www.tidio.com/blog/ai-test

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

Sinapayen, L. (2023). Telling Apart AI and Humans #3: Text and humor https://towardsdatascience.com/telling-apart-ai-and-humans-3-text-and-humor-c13e345f4629

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. (2023). How Far Can We Get in Creating a Digital Replica of a Philosopher? In *Social Robots in Social Institutions* (pp. 371–380). IOS Press. https://doi.org/10.3233/FAIA220637

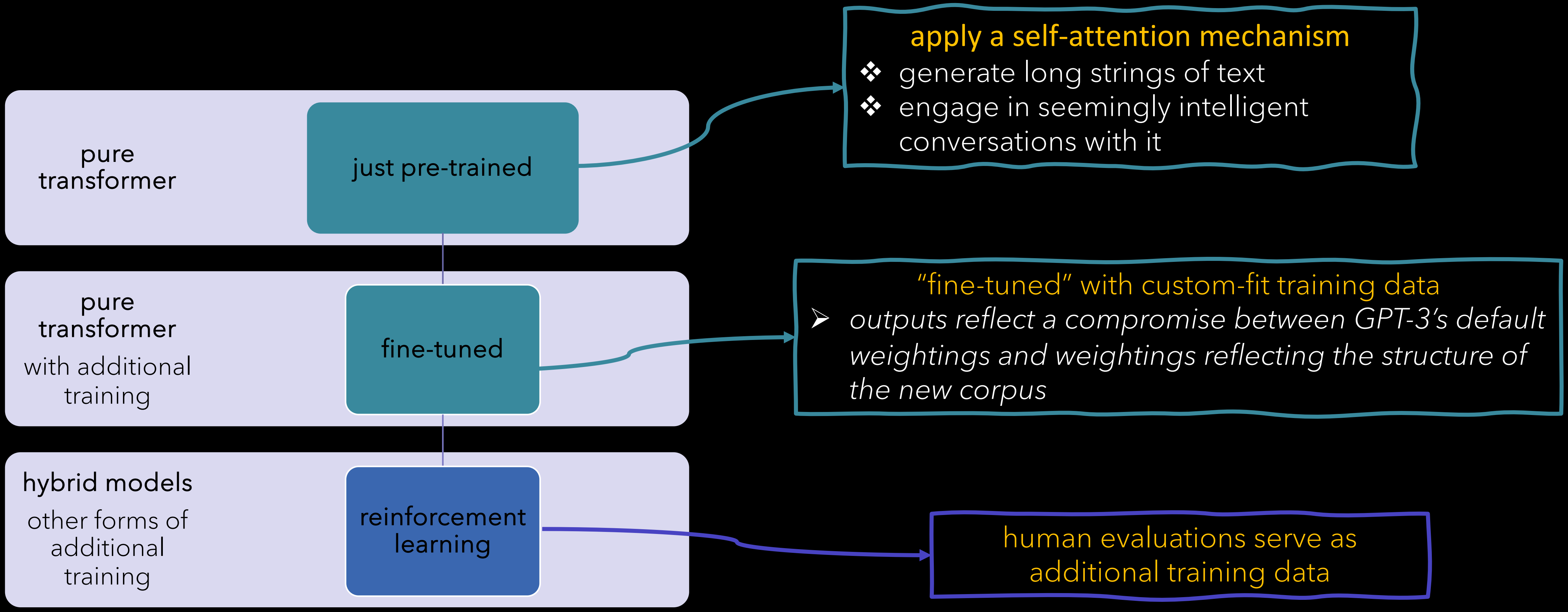
Strasser, Anna (2023). On pitfalls (and advantages) of sophisticated Large Language Models. preprint at https://arxiv.org/abs/2303.17511

Vota, W. (2020). Bot or Not: Can You Tell What is Human or Machine Written Text? https://www.ictworks.org/bot-or-not-human-machine-written/#.Y9VO9hN_oRU

YOV – Build A Versona—Never Have to Say Goodbye. (2023, March 23). https://www.myyov.com/index.html

Wolfram, S. (2023). What Is ChatGPT Doing … and Why Does It Work. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work

OTHER LARGE LANGUAGE MODELS



pure transformer

just pre-trained

apply a self-attention mechanism

- generate long strings of text
- engage in seemingly intelligent conversations with it

pure transformer

with additional training

fine-tuned

"fine-tuned" with custom-fit training data

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

hybrid models

other forms of additional training

reinforcement learning

human evaluations serve as additional training data