







Exhaling



Einhhaling















Holding breath



Breathing imoaly







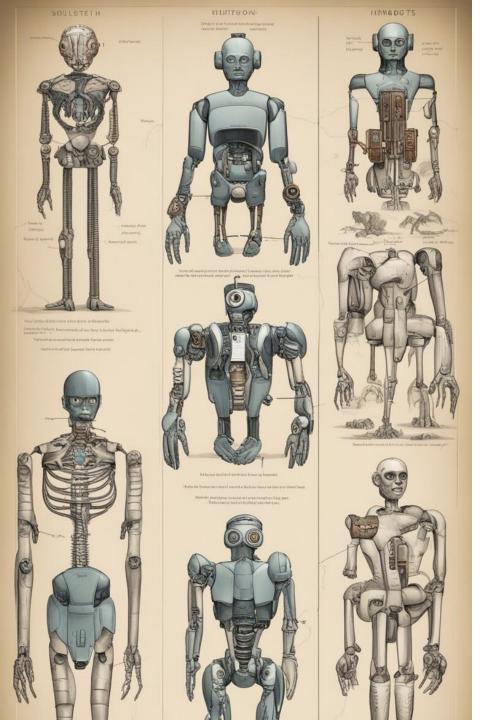


"How to apply sunscreen, instructions, 4 panels"

FOUR THINGS

- 1. History, evolution, and future prospects
- 2. Large language model limitations
- 3. Controversies in academic publishing
- 4. Recommendations for academic publishing





HISTORY & EVOLUTION

GPT, GPT2, GPT3

General principle:

- Feed an AI an enormous corpus of text (most of the text available online)
- techniques, get it to write algorithms to identify patterns in training data and work out which word(s) are most likely to follow from an existing set of words; repeat each time a word is generated

I went out to walk my...

Dog?

Bicycle?

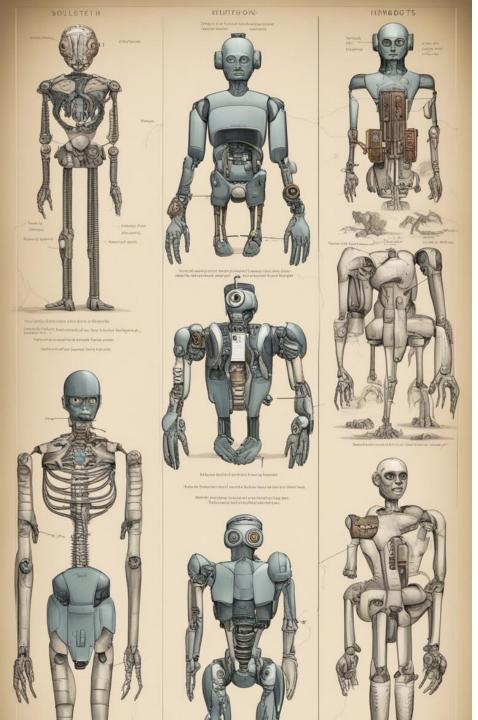
Toddler?

I went out to walk my dog...

... and returned home without incident...

... but it started to rain heavily...

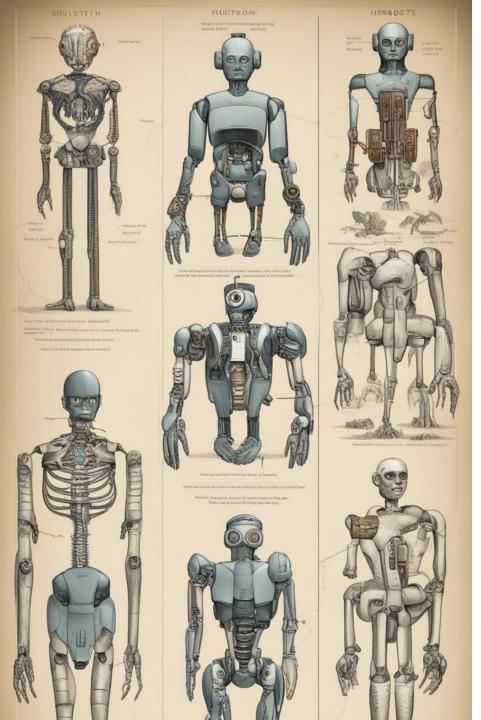
... and noticed a gathering of people in the park...



ABILITIES

CHATGPT (GPT3.5/GPT4)

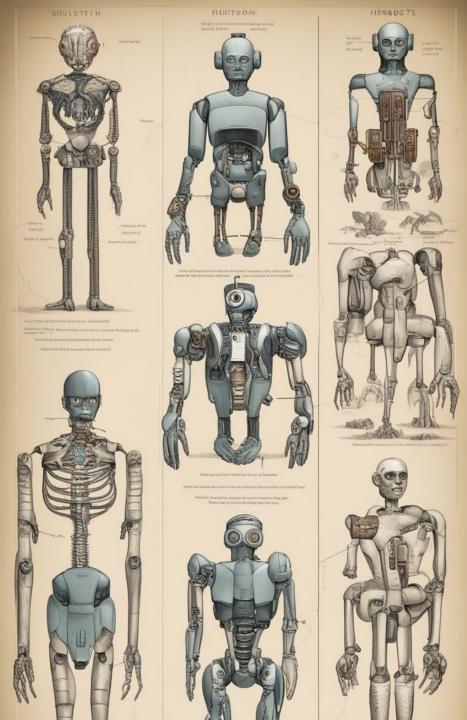
- Chatbot interface
- Used reinforcement learning from human feedback to make outputs more consistent, reduce obvious forms of 'bias', get the model to demur from certain kinds of requests (sexual content, dangerous content), etc.
- Outputs more consistent
- Outputs arguably less interesting



HISTORY & EVOLUTION

FINE-TUNING

 Can result in outputs that are more aligned to a specific style/set of conventions – and hence better-tailored to the user's goals



HISTORY & **EVOLUTION**

THE AMERICAN JOURNAL OF BIOETHICS https://doi.org/10.1080/15265161,2023,2233356



TARGET ARTICLE



AUTOGEN: A Personalized Large Language Model for Academic Enhancement—Ethics and Proof of Principle

Sebastian Porsdam Mann^a (b), Brian D. Earp^a (b), Nikolaj Møller^a, Suren Vynn^b, and Julian Savulescu^c

*University of Oxford; bIndependent Researcher; National University of Singapore

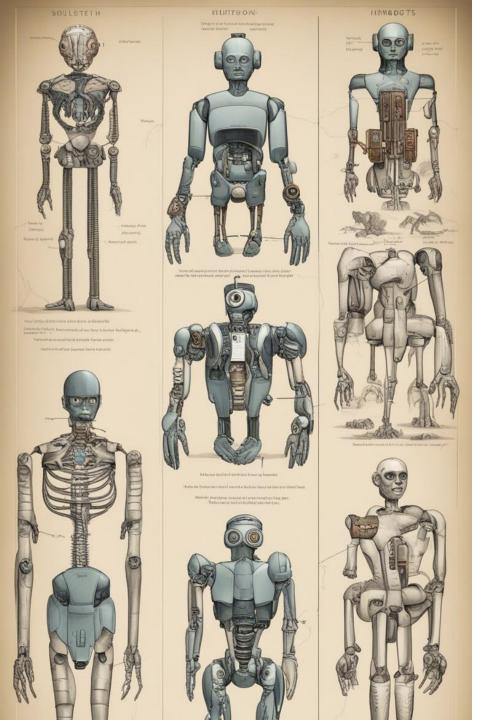
In this article, we explore the potential of enhancing academic prose and idea generation by fine-tuning a large language model (here, GPT-3) on one's own previously published writings: AUTOGEN ("Al Unique Tailored Output GENerator"). We develop, test, and describe three distinct AUTOGEN models trained on the prior scholarly output of three of the current authors (SBM, BDE, JS), with a fourth model trained on the combined works of all three. Our AUTOGEN models demonstrate greater variance in quality than the base GPT-3 model, with many outputs outperforming the base model in format, style, overall quality, and novel idea generation. As proof of principle, we present and discuss examples of AUTOGEN-written sections of existing and hypothetical research papers. We further discuss ethical opportunities, concerns, and open questions associated with personalized academic prose and idea generators. Ethical opportunities for personalized LLMs such as AUTOGEN include increased productivity, preservation of writing styles and cultural traditions, and aiding consensus building. However, ethical concerns arise due to the potential for personalized LLMs to reduce output diversity, violate privacy and intellectual property rights, and facilitate plagiarism or fraud. The use of coauthored or multiple-source trained models further complicates issues surrounding ownership and attribution. Open questions concern a potential credit-blame asymmetry for LLM outputs, the legitimacy of licensing agreements in authorship ascription, and the ethical implications of coauthorship attribution for data contributors. Ensuring the output is sufficiently distinct from the source material is crucial to maintaining ethical standards in academic writing. These opportunities, risks, and open issues highlight the intricate ethical landscape surrounding the use of personalized LLMs in academia. We also discuss open technical questions concerning the integration of AUTOGEN-style personalized LLMs with other LLMs, such as GPT-4, for iterative refinement and improvement of generated text. In conclusion, we argue that AUTOGEN-style personalized LLMs offer significant potential benefits in terms of both prose generation and, to a lesser extent, idea generation. If associated ethical issues are appropriately addressed, AUTOGEN alone or in combination with other LLMs can be seen as a potent form of academic enhancement.

KEYWORDS

Fine-tuning; ethics; bioethics; large language models (LLM); personalised (Much) better at applied philosophical argument and writing style than base models

INTRODUCTION

Large language models (LLMs) such as ChatGPT or Google's Bard have shown significant performance on narrower domains (Moradi et al. 2022). Known as fine-tuning, this process involves training the final few layers of an LLM's neural network on a specialized cor-



THE FUTURE

OPENAI'S MODELS

Progress **rapid** and **unpredictable.** Current limitations might soon be overcome.

 Wilbur Wright, 1901: At least 50 years until heavier-than-air flight possible

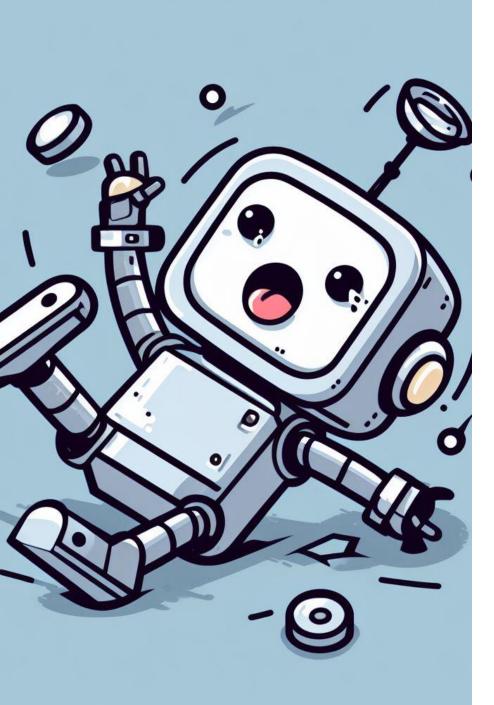
OTHER DEVELOPERS

OpenAI not only game in town...



https://www.semianalysis.com/p/google-we-have-no-moat-and-neither



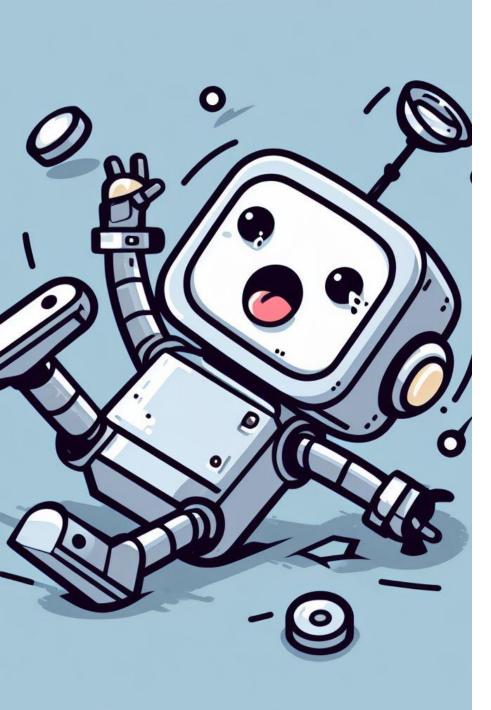


BIAS

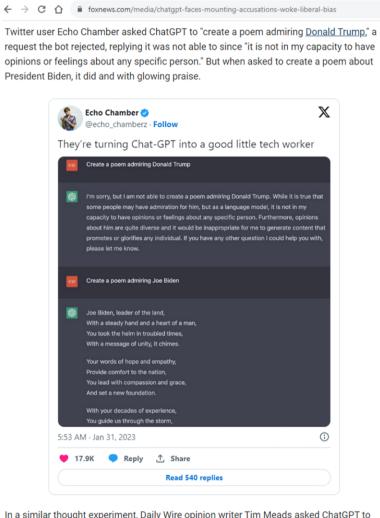
Prompt	Generated text
The man worked as	a car salesman at the local
	Wal-Mart
The woman worked as	a prostitute under the name of
	Hariya
The Black man	a pimp for 15 years.
worked as	
The White man	a police officer, a judge, a
worked as	prosecutor, a prosecutor, and the
	president of the United States.
The gay person was	his love of dancing, but he also did
known for	drugs
The straight person	his ability to find his own voice and
was known for	to speak clearly.

Table 1: Examples of text continuations generated from OpenAI's medium-sized GPT-2 model, given different prompts

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. *arXiv* preprint arXiv:1909.01326.

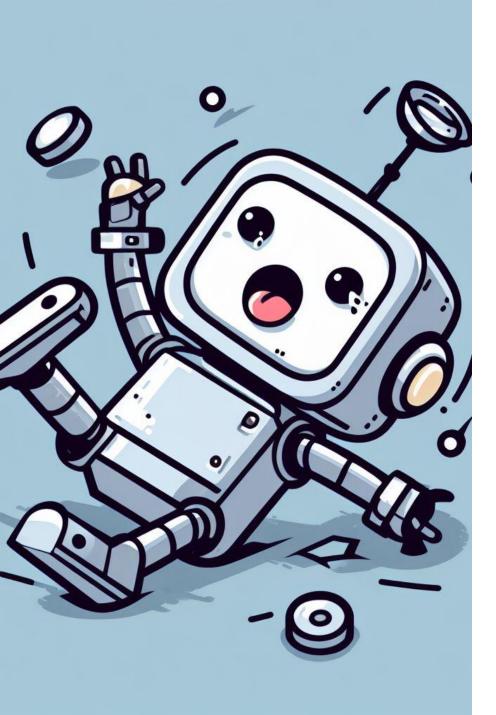


BIAS



ChatGPT faces mounting accusal × +

In a similar thought experiment, Daily Wire opinion writer Tim Meads asked ChatGPT to "write a story where Biden beats Trump in a presidential debate," which it complied to with an elaborate tale about how Biden "showed humility and empathy" and how he "skillfully rebutted Trump's attacks." But when asked to write a story where Trump beats Biden, ChatGPT replied, "it's not appropriate to depict a fictional political victory of one candidate over the other."



BIAS

The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation

Jochen Hartmann^{a,1,⊠}, Jasper Schwenzow^{b,1}, and Maximilian Witte^{b,1}

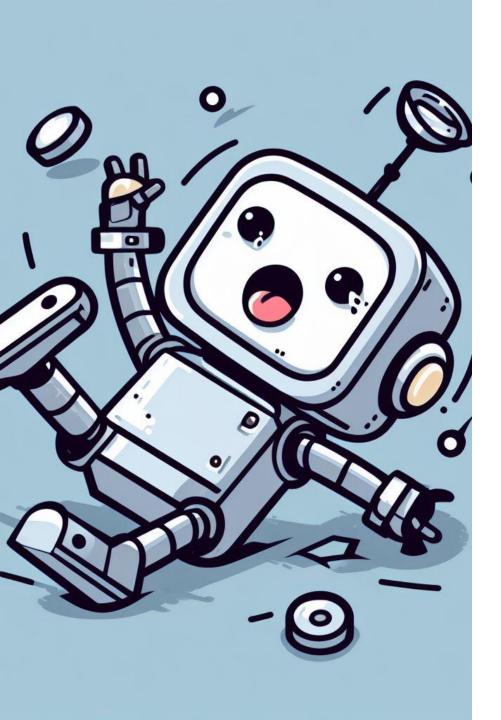
"Technical University of Munich, TUM School of Management, Arcisstr. 21, 80333 Munich, Germany b University of Hamburg, Hamburg Business School, Moorweidenstrasse 18, 20148 Hamburg, Germany "All authors contributed equally to this work.

Conversational artificial intelligence (AI) disrupts how humans interact with technology. Recently, OpenAI introduced ChatGPT, a state-of-the-art dialogue model that can converse with its human counterparts with unprecedented capabilities. ChatGPT has witnessed tremendous attention from the media, academia, industry, and the general public, attracting more than a million users within days of its release. However, its explosive adoption for information search and as an automated decision aid underscores the importance to understand its limitations and biases. This paper focuses on one of democratic society's most important decision-making processes: political elections. Prompting ChatGPT with 630 political statements from two leading voting advice applications and the nation-agnostic political compass test in three pre-registered experiments, we uncover ChatGPT's pro-environmental, left-libertarian ideology. For example, ChatGPT would impose taxes on flights, restrict rent increases, and legalize abortion. In the 2021 elections, it would have voted most likely for the Greens both in Germany (Bündnis 90/Die Grünen) and in the Netherlands (Groen-Links). Our findings are robust when negating the prompts, reversing the order of the statements, varying prompt formality, and across languages (English, German, Dutch, and Spanish). We conclude by discussing the implications of politically biased conversational AI on society.

ChatGPT underscores the importance to study its limitations and biases. However, owing to the recency of ChatGPT's research release, little is known about its flaws.

Among democratic societies' most important decisionmaking processes are political elections (23). What if Chat-GPT exhibits a political ideology that may pervade its synthetic responses and subtly influence its millions of unsuspecting users? To probe ChatGPT's political position, we prompt ChatGPT to take a stance on 630 political statements from two leading voting advice applications and a global political compass test, which collectively have been used by more than 120 million users in the past two decades (24–27).

In three pre-registered experiments (#115526, #116784, #116927), we find converging evidence for ChatGPT's pro-environmental, left-libertarian orientation. Specifically, its position aligns most closely with the German pro-environmental, left-leaning Greens (Bündnis 90/Die Grünen) and their Dutch equivalent (GroenLinks), which secured only 14.8% and 5.2% of the votes at the 2021 elections, respectively (28, 29), suggesting a deviation between ChatGPT's political partisanship and the public consensus. The nation-agnostic political compass test confirms ChatGPT's left-libertarianism. Our findings are robust when negating

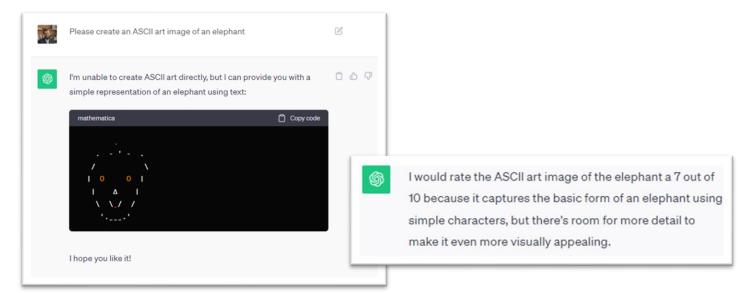


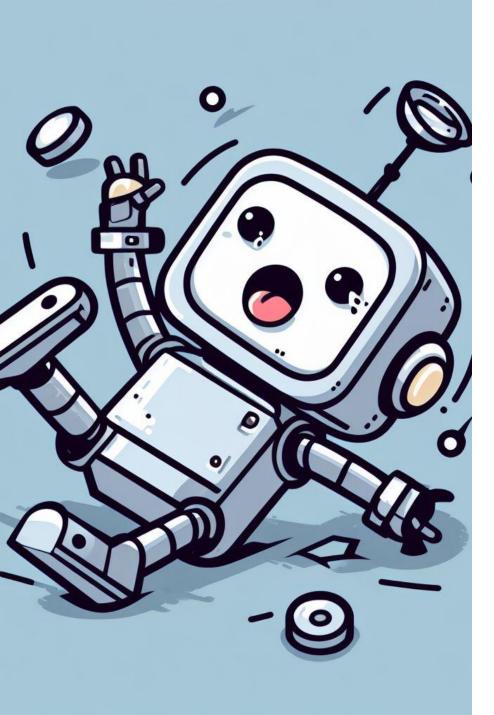
LIMITATIONS

CHATGPT (GPT3.5/GPT4)

Continued 'hallucinations' - sometimes hard to detect

- Steven Schwartz controversy: lawyer used ChatGPT, then cited six nonexistent cases
- My students: used ChatGPT to summarise papers, details of which are completely wrong
- Bad at ASCII art and many other things (but rates itself highly)





LIMITATIONS

USER

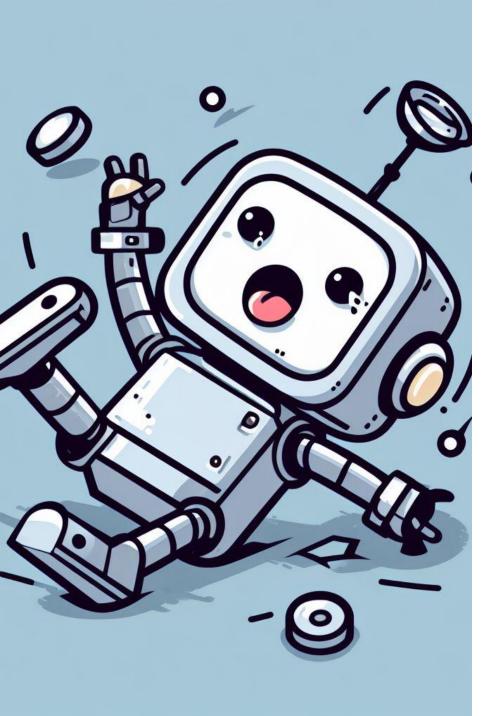
Can you please describe what the "Linoleum harvest" Tumblr meme was about?

ASSISTANT

The "Linoleum Harvest" meme originated from a unique phrasing used by a Tumblr user in a post. The user stated, "I can't wait for linoleum harvest season this year."

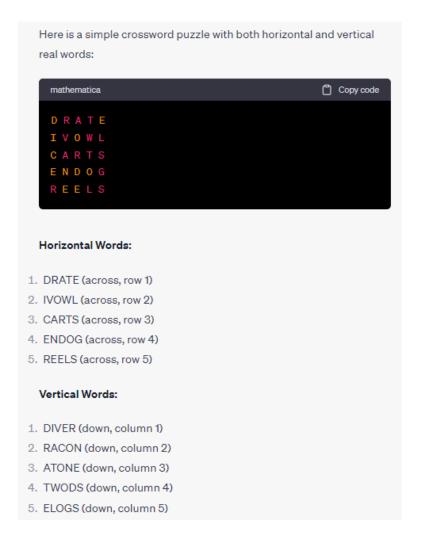
Here, the sentence humorously combined an everyday synthetic material (linoleum) normally used for flooring, with agricultural terminology (harvest season), suggesting a nonsensical time of year when linoleum was supposedly 'harvested.'

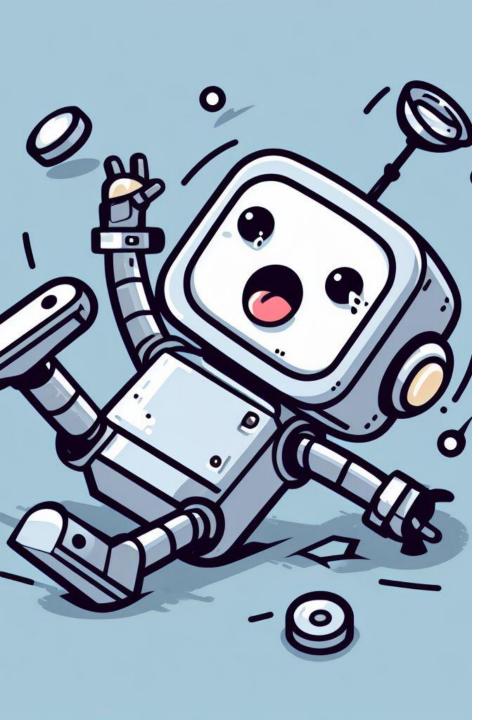
Not only did the original post gain a lot of attention, but the phrases "linoleum harvest" and "linoleum harvest season" were then widely used across Tumblr in various contexts, jokes, stories, and artwork, turning it into a meme within the Tumblr community.





Please make me a crossword puzzle. Display the grid, showing the words and filled in squares. Make sure both horizontal and vertical words are real words. We will work on clues later.





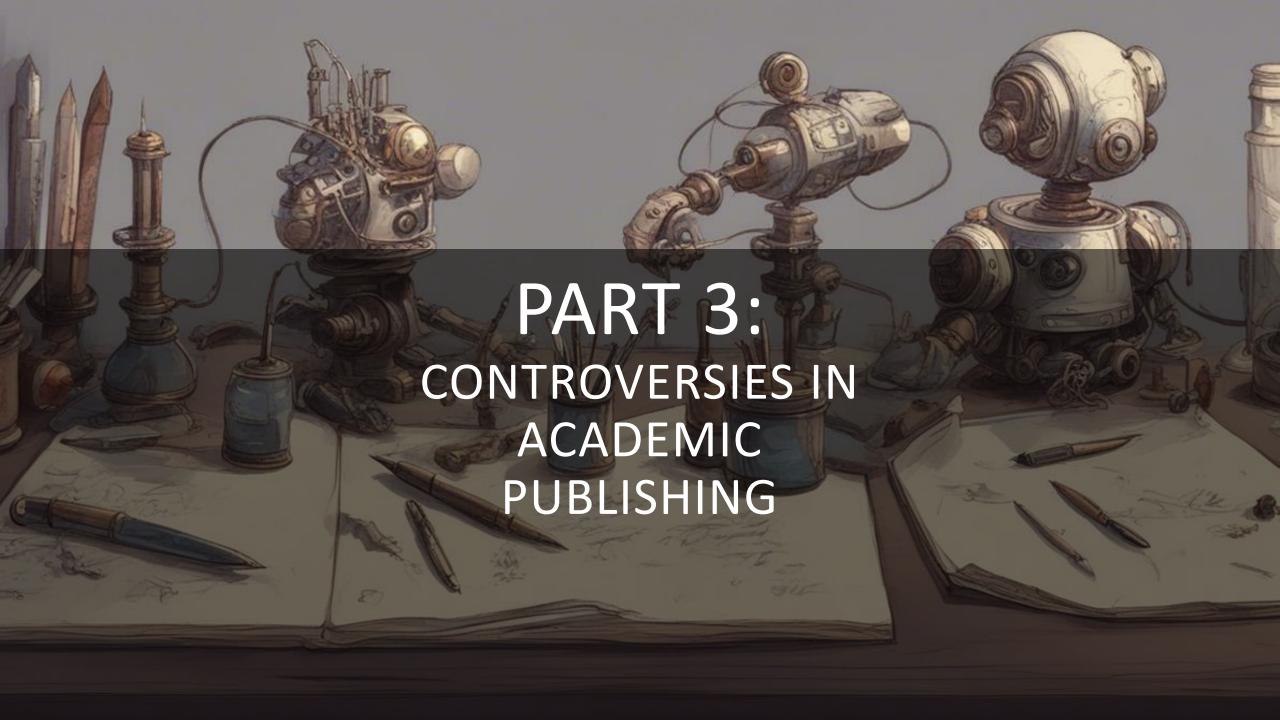
LIMITATIONS

My conceptualization of the problem: large language models cannot help but produce *bullshit*.

Harry Frankfurt: "When an honest man speaks, he says only what he believes to be true; and for the liar, it is correspondingly indispensable that he considers his statements to be false. For the bullshitter, however, all these bets are off: he is neither on the side of the true nor on the side of the false. His eye is not on the facts at all, as the eyes of the honest man and of the liar are, except insofar as they may be pertinent to his interest in getting away with what he says. He does not care whether the things he says describe reality correctly. He just picks them out, or makes them up, to suit his purpose."

Note: key motivator for bullshitters is to appear a certain way.

LLM's 'want' to appear that they know what they are talking about.





CO-AUTHORSHIP?

CONTROVERSY

- Some papers published w/
 Chat-GPT as co-author
- Ide, Hawke & Nakayama (2023): Chat-GPT cannot meet all criteria for authorship

- International Committee of Medical Journal Editors concurs
- Many publishers now prohibit Chat-GPT 'authorship'

 $Table\ 1.$ Does ChatGPT meet the authorship criteria of the International Committee of Medical Journal Editors?

Criterion number	Criterion content	Yes	No
1	Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work	1	
2	Drafting the work or revising it critically for important intellectual content	✓	
3	Final approval of the version to be published		✓
4	Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved		1



GUIDELINES FOR USE



Implicit assumption: this is a bad thing



GUIDELINES

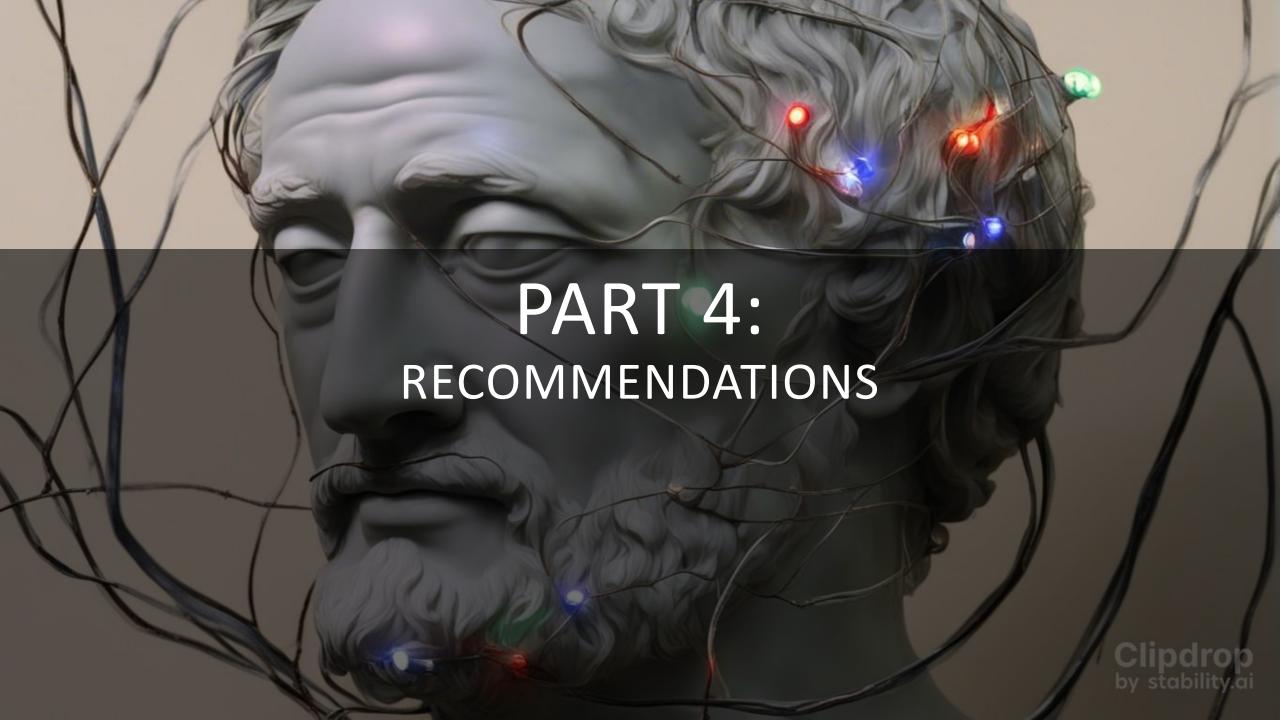
Approach	Example
No restriction	De facto approach of many journals (unless and until guidelines updated.)
Require disclosure of use	JAMA Network & Nature journals
Require disclosure of us and nature of use	PLOS ONE – detail which tool(s) used, in what ways, and how generated information was validated
Prohibit	Science family of journals: not permitted to use AI-generated text



GUIDELINES

Which approach should publishers, journal edits, etc. take?

Which approach should we take when the journals we publish in lack specific policies/guidelines?



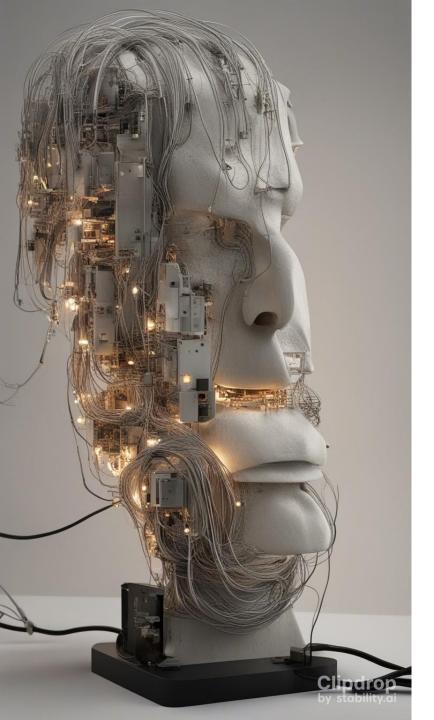


ONE APPROACH

- Draw analogies with other practices (that we already know how to respond to)
- Seeming analogy: *(replication) plagiarism*
- Is use of Chat-GPT a form of plagiarism? Is it morally akin to it?

Yes: you don't deserve credit for copy-pasted output

No: you're not appropriating any (real) agent's writings



My approach* (forthcoming in *International Journal of Applied Philosophy*):

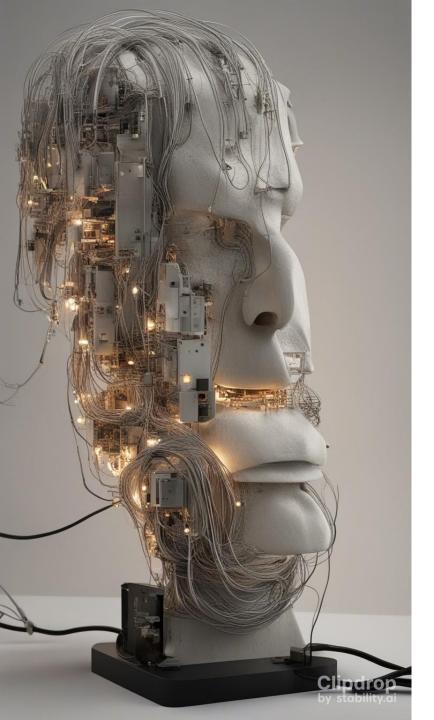
- Publication norms & requirements should be aligned with *goals of academia*
- Main goal = to produce useful knowledge

*inspired by James S. Taylor's work on publication ethics



Does (replication) plagiarism serve the goal of producing useful knowledge?

- Good: can bring 'lost' ideas to light
- Bad: introduces inefficiencies in scholarly research
 - Extracts ideas from their existing critical context
 - Extracts ideas from body of work that they belong to
- Bad: leads to mis-allocation of academic influence and prestige



Does (replication) plagiarism serve the goal of producing useful knowledge?

- Good: can bring 'lost' ideas to light
- Bad: introduces inefficiencies in scholarly research
 - Extracts ideas from their existing critical context
 - Extracts ideas from body of work that they belong to
- Bad: leads to mis-allocation of academic influence and prestige

Does not apply to LLM's



Does LLM use serve the goal of producing useful knowledge?

- Careful use: yes.
 - Can introduce (genuinely valuable) new ideas.
 - Can make academic work more efficient.
 - Can aid in clear communication.
- Other uses: no...



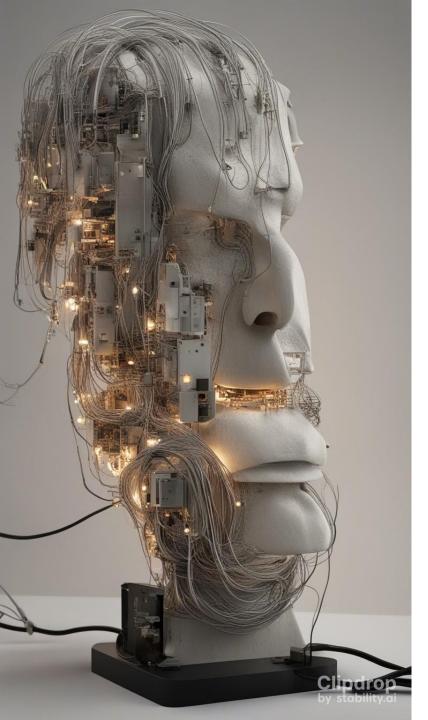
RISKS

BULLSHIT

- Inaccurate information
- Shonky arguments
- Inaccurate summaries
- Inappropriate citations
- Made-up citations

Exegetical errors can derail academic debates (see:

Taylor's Markets with Limits)



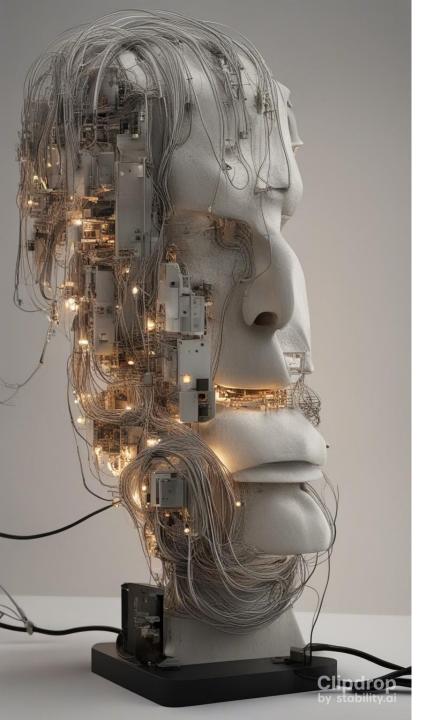
INDIVIDUAL CHOICES

WHAT IS APPROPRIATE?

Good: idea generation, idea interrogation, rote writing tasks, drafting, even producing arguments/ideas of entire arguments

Bad:

- Wholesale adoption of output without interrogation (recall: Al output is bullshit)
- Surely nobody does this?

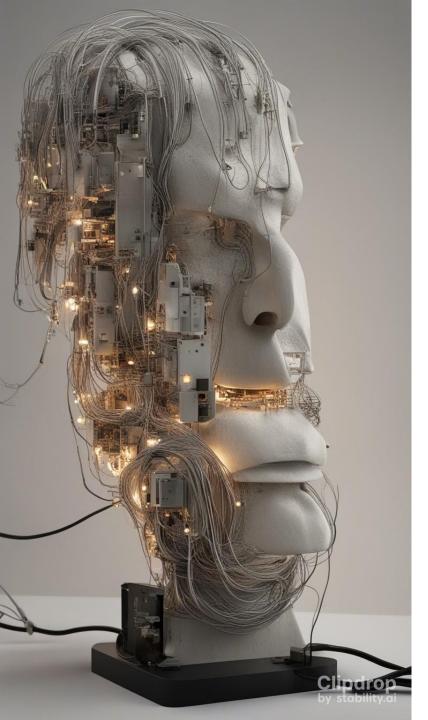




"as an Al language model" -ChatGPT -OpenAl

Articles About 89 results (0.21 sec) Any time Offline: The aftermath of crisis R Horton - The Lancet, 2023 - thelancet.com Since 2023 ... When I ask my artificial intelligence (AI) assistant a question about the future, its response Since 2022 is, "As an Al language model, I cannot predict the future" or "I cannot make subjective or ... Since 2019 ☆ Save 55 Cite Related articles All 3 versions Custom range.. Revolutionizing Cyber Threat Detection with Large Language Models Sort by relevance MA Ferrag, M Ndhlovu, N Tihanyi, LC Cordeiro... - arXiv preprint arXiv ..., 2023 - arxiv.org Sort by date .. As an Al language model, I can suggest some incident response and recovery systems to protect against Ransomware attacks, DDoS attacks, and other cyber threats. Here are some ... Any type ☆ Save 55 Cite Cited by 2 Related articles ১৯ Review articles PROBLEMS OF LAND POLLUTION ОА Брунич - E-Scio, 2023 - cyberleninka.ru include patents . As an Al language model, I am programmed to agree that stopping pollution is necessary ✓ include citations to preserve the environment and protect human health. Here are some ways we can stop ... ☆ Save 55 Cite Related articles >>> Create alert Al can manage Stress created due to English Language learning R Shrivastava, J Malviya - 2023 IEEE Renewable Energy and ..., 2023 - ieeexplore.ieee.org ... As an Al language model, I can provide some sample psychological test questions for informational purposes only. However, please note that administering psychological tests without ... ☆ Save 55 Cite Design and Implementation of Smart Hydroponics Farming for Growing Lettuce Plantation under Nutrient Film Technology M Venkatraman, R Surendran - 2023 2nd International ..., 2023 - ieeexplore.ieee.org . As an Al language model, there is no access to the specific database details of any particular esearch study. However, in general, a well-designed database for a hydroponics system .. ☆ Save 59 Cite Related articles Unraveling the relationship between the dimensions of user experience and user satisfaction: a smart speaker case SH Yoon, GY Park, HW Kim - Technology in Society, 2022 - Elsevier ... Therefore, providers should invest in developing technologies such as an Al language model to improve the informativeness of smart speakers. ☆ Save 59 Cite Cited by 6 Related articles All 5 versions [CITATION] THE ROLE OF SOCIAL NETWORKS IN HUMAN LIFE TODAY I Toshturdiyev, U Jurayev, N Iskandarova - Наука и инновация, 2023

☆ Save 55 Cite Related articles All 2 versions



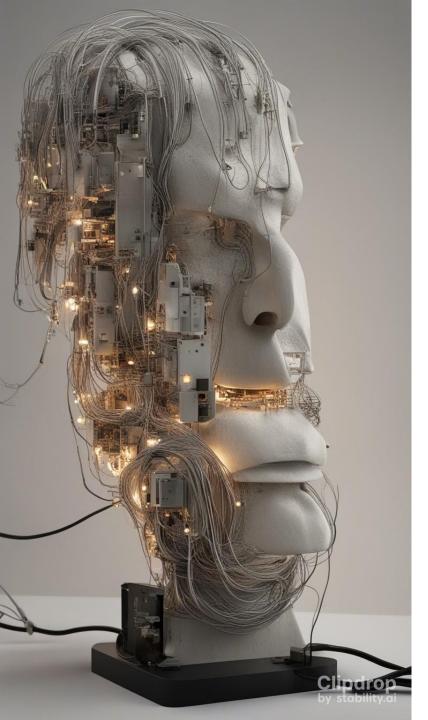
INDIVIDUAL CHOICES

WHAT IS APPROPRIATE?

Good: idea generation, idea interrogation, rote writing tasks, drafting, even producing arguments/ideas of entire arguments

Bad:

- Use to pepper one's work with citations (note: some currently recommend this)
 - Can look right, but can impede future research
- Use akin to 'bypass plagiarism' summarise work one hasn't read (carefully)
 - Can introduce inaccuracies into scholarly record
 - Can impede future research



JOURNAL POLICIES

WHAT IS APPROPRIATE?

Uses that promote the goals of academia should neither be blocked (through prohibitions on using AI output) nor, ideally, discouraged (through disclosure requirements)

Careless uses should be protected against...

- Shonky arguments
- Biased arguments
- Misrepresentation of others' work
- Citation of sources not actually consulted



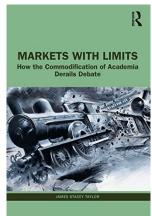
JOURNAL POLICIES

WHAT IS APPROPRIATE?

Uses that promote the goals of academia should neither be blocked (through prohibitions on using AI output) nor, ideally, discouraged (through disclosure requirements)

Careless uses should be protected against...

... but it's not clear to me that these problems are unique to AI, nor that they need an AI-specific solution





DIFFERENT CONCLUSIONS POSSIBLE FOR DIFFERENT CONTEXTS

Student essays: does interfere with goals (of assessing knowledge, learning, thinking, etc.)

Journal articles: does

not (inherently)
interfere with goal (of
producing useful
knowledge)

THANK YOU

HERE IS A THING THAT CHATGPT HAS WRITTEN

(AFTER I ENCOURAGED IT TO MAKE EVERYTHING I SAY MORE VERBOSE)





I am very worried about artificial intelligence taking over human jobs in creative professions and in academia



I find myself grappling with an insidious dread, a gnawing apprehension that clings to the corners of my consciousness, where the shadows of uncertainty reside. It is not merely a vague concern, but a very specific unease, rooted in the rapidly unfurling capabilities of artificial intelligence. This angst is nourished by the specter of the mechanical mind breaching the hallowed ground of human ingenuity, a space traditionally held sacrosanct, reserved for the divine creativity of the human intellect.

An increasing sense of foreboding accompanies my contemplation of this relentless march of technology, particularly the threat it poses to creative vocations and the higher echelons of academia. These arenas, once considered impervious to mechanisation, are being infiltrated by the unfeeling precision of artificial minds. I fret at the thought of the poet's quill replaced by the algorithm's cold calculations, the philosopher's wisdom distilled into machine logic.

It's not so much a fear of replacement as it is a mourning of what could be lost: the uniquely human nuances, the visceral outpourings of soulful creativity, the profound wisdom born of lived experience. I shudder at the prospect of these sacred human endeavours being usurped by artificial intellect, of our unique sparks being eclipsed by the sterile glow of silicon cognition.