Generative AI and research integrity

GPT-based text generators like ChatGPT or Microsoft Copilot have rapidly become a "cultural sensation" [1]. This document provides scholars with guidelines and scientific background on how to think critically and mindfully about these tools in academic writing and research.

Preamble

All research at our institution, from ideation and execution to analysis and reporting, is bound by the Dutch Code for Research Integrity. This code specifies five core virtues that organise and inform research conduct: Honesty, Scrupulousness, Transparency, Independence and Responsibility.

One way to summarise the guidelines in this document is to say they are about taking these core virtues seriously. When it comes to using Generative AI in or for research, the question is if and how this can be done honestly, scrupulously, transparently, independently, and responsibly.

A key ethical challenge is that most current Generative AI undermines these virtues by design [3–5; details below]. Input data is legally questionable; output reproduces biases and erases authorship; fine-tuning involves exploitation; access is gated; versioning is opaque; and use taxes the environment.

While most of these issues apply across societal spheres, there is something especially pernicious about text generators in academia, where writing is not merely an output format but a means of thinking, crediting, arguing, and structuring thoughts. Hollowing out these skills carries foundational risks.

A common argument for Generative AI is a promise of higher productivity [5]. Yet productivity does not equal insight, and when kept unchecked it may hinder innovation and creativity [6, 7]. We do not need more papers, faster; we rather need more thoughtful, deep work, also known as slow science [8–10].

For these reasons, the first principle when it comes to Generative AI is to not use it unless you can do so honestly, scrupulously, transparently, independently and responsibly. The ubiquity of tools like ChatGPT is no reason to skimp on standards of research integrity; if anything, it requires more vigilance.

Scientific background: What GenAI tools are and what they are not

What they are: Generative AI tools are statistical models fitted to enormous training datasets that can produce output along the lines of statistical patterns found in the data [3, 11]. Because they rely on probabilistic prediction instead of reasoning and understanding, their output (in relation to a prompt) should never be treated as trustworthy or reliable, though it may seem (for topics well-represented in training data) plausible and predictable [12]. Because training data is often of questionable legality, much of Generative AI output is problematic both ethically (erasing credit and authorship attribution) and legally (violating original licenses) [13, 14]. Because these tools are computationally demanding, their carbon footprint (in training and in use) is considerable [15].

What they are not: Despite appearances, Generative AI tools are not 'creative' or 'intelligent' in the sense of being able to reason or understand [16, 17], and they cannot be considered authors (of texts) or artists (of images) [1]. Because they easily reproduce or amplify statistical biases in the training data, they are neither original nor objective [3]. Because they generate output probabilistically, they are fundamentally unlike search engines or information retrieval systems [18]. And finally, because their output is stochastic and not well-defined nor well-scoped [19], they are fundamentally unlike calculators [20, 21]. A common rhetorical frame is that just as pocket calculators replaced slide rules, so Generative AI will replace… what exactly? Working out the ways this analogy misrepresents the realities and complexities of scholarly research and writing is an exercise left to the reader.

What they do: Generative AI tools produce superficially plausible output by regurgitating the most likely statistical patterns from the training data given an input prompt. They excel in variations on a theme, especially when that theme is well-represented in training data (e.g., Wikipedia-like factoids, books and images of certain genres or styles, programming code and explanations). Reflecting this, scholarly experts have formulated a number of deflationary ways to think about Generative AI: stochastic parrots [3]; spicy autocomplete [22]; and ways of automating plagiarism [23]. Though partial, these similes do more justice to the nature of these tools than the imprecise buzzword 'artificial intelligence' [24, 25].

How they are designed: Large Language Models like ChatGPT are designed as interactive interfaces that enter into a "chat" with the user. Using human labour (in the case of OpenAI, exploiting Kenyan workers [26]), their output is tuned to come across as helpful, confident, chatty and inoffensive. Because of this, they will produce statements that sound plausible (but lack understanding); they will 'apologise' when corrected (but fail to learn); and they will evade some types of output (but remain vulnerable to workarounds). Also because of this, naïve users are likely to anthropomorphise them, overattributing agency and autonomy [27] and underestimating their own suggestibility [28]. Understanding the 'dark patterns' [29] of the design of Generative AI tools is crucial for grasping their persuasiveness and for explaining people's susceptibility to their output.

Principles

Writing

- In many cases, it is a bad idea to use Generative AI in academic writing. Reasons include:
    - By design, it rehashes text without understanding, authorship attribution, or credit: a form of plagiarism [30]. Honest, scrupulous, and transparent scholarly work requires the opposite.
    - By design, it excels at regurgitating the most plausible-sounding 'average' take on a subject. Independent, responsible and innovative scholarly work requires the opposite [31, 32].
    - By design, it makes producing texts seem deceptively easy. But for scholars, writing —wrestling with words— is thinking, and to outsource writing is to give up on thinking [33].
- In some limited and well-circumscribed cases, Generative AI may be useful. This may include:
    - Looking up synonyms or alternative wordings; generating multiple variations of a text you have authored to learn how to write better; or brainstorming project names.
    - However, consider that word processors have a thesaurus; reading widely will expose you to stylistic variation alongside new ideas; boilerplate text rarely makes for engaging writing; and your colleagues are often happy to provide feedback and brainstorm with you.

Research

- There are serious risks and downsides to using Generative AI in research. These include:
    - Noise, biases, and spurious features of prompt design make results less transparent [34–36]
    - Black-boxed proprietary systems make analyses unstable and hamper reproducibility [37]
    - Loss of human judgement in analytical choices dilutes responsibility and independence [38, 39]
- In some limited and well-circumscribed cases, Generative AI may be useful. Examples include:

- Correcting or cleaning up OCR errors in scanned materials, e.g., historical sources [5].
- Generating synthetic data for testing analyses, or for data privacy & protection [40].
- Augmenting datasets with summaries or other forms of metadata [36].
- However, consider that for many such uses, established analysis pipelines exist that offer more transparency, reproducibility and independence [41, 42].

Writing & research

If you do consider using Generative AI in writing or research, follow these principles:

- Choose maximally open, local, non-proprietary models to ensure transparency, reproducibility and independence [37, 43].
- Think carefully about the consequences of outsourcing data annotation, classification, summarisation or augmentation to a stochastic model — and take full responsibility for errors and unforeseen uses. Think ahead about how you will prevent abuse and justify your choices to peer reviewers.
- Never share personal, privacy-sensitive or work-related data outside your organisation, including through prompts or automated APIs of proprietary services. There are multiple documented instances of such data (including draft research proposals) being leaked [44]. Think ahead about how you would justify your choices in response to an ethics review board or a GDPR access request.
- Always disclose use of Generative AI, and be technically precise and responsible in doing so. This includes clearly identifying synthetically generated media and avoiding anthropomorphic descriptions [27, 45].
- When generating output for research purposes, scrupulously keep track of prompts, model versions, and dates, using established and reproducible computational workflows [46].
- Never pass off Generative AI output as human-generated. In scientific writing, this counts as misconduct, and failure to disclose has already led to retractions. In research, there is a possible exception: when carrying out research on Generative AI, disclosure may not always be desired. Responsible disclosure should still happen in the ethics review procedure and in reporting results.

About these guidelines

Complex matters rarely have simple answers. Generative AI tools pose novel dilemmas and highlight the need to build critical AI literacy [47]. We encourage you to engage in conversation about these matters with colleagues, and to call on expertise where needed.

These guidelines were drawn up at the request of the directorates of the Centre for Language Studies (CLS) and Radboud Institute for Culture and History (RICH) at the Faculty of Arts, Radboud University, Nijmegen. Written by Mark Dingemanse and revised based on feedback from Andreas Liesenfeld, Ada Lopez, Tamar Sharon, Iris van Rooij, Olivia Guest, Liedeke Plate, Enny Das, Gijske de Boo, and other interlocutors online and offline.

This document will be revised with new information as the generative AI landscape continues to evolve. We encourage you to always consult the most recent version. This is version 1, April 2024, reflecting the state of technology and scholarly literature in Spring 2024.

Bibliography

Key readings

- Goodlad, Lauren M.E., Sharon Stoerger, and AI Round Table Advisory Council. 2024. 'Rutgers AI Council: Teaching Critical AI Literacies'. https://otear.rutgers.edu/initiatives/ai/ai-advice/
- Messeri, Lisa, and M. J. Crockett. 2024. 'Artificial Intelligence and Illusions of Understanding in Scientific Research'. *Nature* 627 (8002): 49–58. https://doi.org/10.1038/s41586-024-07146-0.
- Widder, David Gray, Sarah West, and Meredith Whittaker. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. https://doi.org/10.2139/ssrn.4543807

References cited

1. Thorp, H. Holden. 2023. ChatGPT is fun, but not an author. *Science* 379: 313–313. doi: 10.1126/science.adg7879.
2. O'Neil, Cathy. 2017. Weapons of math destruction: how big data increases inequality and threatens democracy. New York: B/D/W/Y Broadway Books.
3. Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. Virtual Event Canada: ACM. doi: 10.1145/3442188.3445922.
4. McQuillan, Dan. 2022. *Resisting AI: an anti-fascist approach to artificial intelligence*. Bristol, UK: Bristol University Press.
5. Karjus, Andres. 2023. Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence. arXiv. doi: 10.48550/arXiv.2309.14379.
6. Chu, Johan S. G., and James A. Evans. 2021. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences* 118. National Academy of Sciences. doi: 10.1073/pnas.2021636118.
7. Ullberg, Eskil. 2015. Productivity Versus Creativity. In *New Perspectives on Internationalization and Competitiveness: Integrating Economics, Innovation and Higher Education*, ed. Eskil Ullberg, 73–91. Cham: Springer International Publishing. doi: 10.1007/978-3-319-11979-3_6.
8. Stengers, Isabelle. 2016. "Another Science Is Possible!": A Plea for Slow Science. In *Demo(s): Philosophy, Pedagogy, Politics*, ed. Hugo Letiche, Geoffrey Lightfoot, and Jean-Luc Moriceau, 53–70. Brill. doi: 10.1163/9789462096448_004.
9. Frith, Uta. 2019. Fast Lane to Slow Science. *Trends in Cognitive Sciences* 0. doi: 10.1016/j.tics.2019.10.007.
10. Alleva, Lisa. 2006. Taking time to savour the rewards of slow science. *Nature* 443: 271–271. doi: 10.1038/443271e.
11. Kockelman, Paul. 2020. The Epistemic and Performative Dynamics of Machine Learning Praxis. *Signs and Society* 8: 319–355. doi: 10.1086/708249.
12. Kocoń, Jan, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* 99: 101861. doi: 10.1016/j.inffus.2023.101861.
13. Samuelson, Pamela. 2023. Generative AI meets copyright. *Science* 381. American Association for the Advancement of Science: 158–161. doi: 10.1126/science.adi0656.
14. Lucchi, Nicola. 2023. ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. *European Journal of Risk Regulation*: 1–23. doi: 10.1017/err.2023.59.
15. Chien, Andrew A, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, 1–7. HotCarbon '23. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3604930.3605705.
16. Bender, Emily M., and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, 5185–5198. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463.
17. Wolfram, Stephen. 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/.
18. Shah, Chirag, and Emily M. Bender. 2024. Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web? *ACM Transactions on the Web* 18: 33:1-33:24. doi: 10.1145/3649468.
19. Bucci, Anthony. 2024. Word calculators don't add up. *Anthony Bucci*. https://bucci.onl/notes/Word-calculators-dont-add-up.
20. Scheider, Simon, Harm Bartholomeus, and Judith Verstegen. 2023. ChatGPT is not a pocket calculator -- Problems of AI-chatbots for teaching Geography. arXiv. doi: 10.48550/arXiv.2307.03196.
21. Ko, Amy J. 2024. More than calculators: Why large language models threaten public education. *Bits and Behavior*. https://medium.com/bits-and-behavior/more-than-calculators-why-large-language-models-threaten-public-education-480dd5300939.
22. Groß, Richard. 2024. Probabilistische Wirklichkeitsmodelle und soziologische Intelligenz. *Soziologie* 53: 60–75.
23. van Rooij, Iris. 2022. Against automated plagiarism. https://irisvanrooijcogsci.com/2022/12/29/against-automated-plagiarism/.
24. van Rooij, Iris, Olivia Guest, Federico G. Adolfi, Ronald de Haan, Antonina Kolokolova, and Patricia Rich. 2023. Reclaiming AI as a theoretical tool for cognitive science. PsyArXiv. doi: 10.31234/osf.io/4cbuv.
25. Mitchell, Melanie. 2019. *Artificial intelligence: a guide for thinking humans*. New York: Farrar, Straus and Giroux.
26. Floridi, Luciano. 2023. AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology* 36: 15. doi: 10.1007/s13347-023-00621-y.
27. Brooker, Phillip, William Dutton, and Michael Mair. 2019. The new ghosts in the machine: "Pragmatist" AI and the conceptual perils of anthropomorphic description. *Ethnographic Studies* 16: 272–298.
28. Zeitlyn, David. 1990. Professor Garfinkel Visits the Soothsayers: Ethnomethodology and Mambila Divination. *Man* 25. New Series: 654–666.
29. Shamsudhin, Naveen, and Fabrice Jotterand. 2021. Social Robots and Dark Patterns: Where Does Persuasion End and Deception Begin? In *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*, ed. Fabrice Jotterand and Marcello Ienca, 89–110. Advances in Neuroethics. Cham: Springer International Publishing. doi: 10.1007/978-3-030-74188-4_7.
30. van Rooij, Iris. 2022. Against automated plagiarism. https://irisvanrooijcogsci.com/2022/12/29/against-automated-plagiarism/.
31. Alon, Uri. 2009. How To Choose a Good Scientific Problem. *Molecular Cell* 35: 726–728. doi: 10.1016/j.molcel.2009.09.013.
32. Lin, Yiling, James A. Evans, and Lingfei Wu. 2022. New directions in science emerge from disconnection and discord. *Journal of Informetrics* 16: 101234. doi: 10.1016/j.joi.2021.101234.
33. Hofstadter, Douglas. 2023. Generative AI Should Not Replace Thinking at My University. *The Atlantic*.
34. Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. arXiv. doi: 10.48550/arXiv.2310.11324.
35. Jacobs, Cassandra. 2023. Why ChatGPT is bad for open psycholinguistics. Substack newsletter. *Scidentity Crisis*. https://cxjacobs.substack.com/p/why-chatgpt-is-bad-for-open-psycholinguistics.
36. Törnberg, Petter. 2024. Best Practices for Text Annotation with Large Language Models. arXiv. doi: 10.48550/arXiv.2402.05129.
37. Liesenfeld, Andreas, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of CUI'23*. Eindhoven.
38. Hong, Sun-ha. 2023. Prediction as extraction of discretion. *Big Data & Society* 10. SAGE Publications Ltd: 20539517231171053. doi: 10.1177/20539517231171053.
39. Crockett, Molly, and Lisa Messeri. 2023. Should large language models replace human participants? OSF. doi: 10.31234/osf.io/4zdx9.
40. James, Stefanie, Chris Harbron, Janice Branson, and Mimmi Sundler. 2021. Synthetic data use: exploring use cases to optimise data utility. *Discover Artificial Intelligence* 1: 15. doi: 10.1007/s44163-021-00016-y.
41. Amrhein, Chantal, and Simon Clematide. 2018. Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods. *Journal for Language Technology and Computational*

Linguistics (JLCL) 33. Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL): 49–76. doi: 10.5167/uzh-162394.

42. Mannino, Miro, and Azza Abouzied. 2019. Is this Real? Generating Synthetic Data that Looks Real. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 549–561. UIST ’19. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3332165.3347866.
43. Widder, David Gray, Sarah West, and Meredith Whittaker. 2023. Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. SSRN Scholarly Paper. Rochester, NY. doi: 10.2139/ssrn.4543807.
44. Krietzberg, Ian. 2024. ChatGPT is leaking users’ passwords, report finds. *TheStreet*. https://www.thestreet.com/technology/chatgpt-sam-altman-artificial-intelligence-privacy-ethics-passwords.
45. Pilling, Franziska, and Paul Coulton. 2019. Forget the Singularity, its mundane artificial intelligence that should be our immediate concern. *The Design Journal* 22: 1135–1146. doi: 10.1080/14606925.2019.1594979.
46. Patel, Ajay, Colin Raffel, and Chris Callison-Burch. 2024. DataDreamer: A Tool for Synthetic Data Generation and Reproducible LLM Workflows. arXiv. doi: 10.48550/arXiv.2402.10379.
47. Goodlad, Lauren M.E., Sharon Stoerger, and AI Round Table Advisory Council. 2024. Rutgers AI Council: Teaching Critical AI Literacies. https://otear.rutgers.edu/initiatives/ai/ai-advice/.