AI AND SCHOLARLY PUBLISHING (HSS)

—

STATUS QUO AND OUTLOOK

GIVEN THE FAR-REACHING ETHICAL IMPLICATIONS OF AI USE IN SCHOLARLY PUBLISHING, WHAT FACTORS SHOULD PUBLISHERS ESPECIALLY CONSIDER WHEN DRAFTING AI GUIDELINES FOR AUTHORS AND EDITORS?

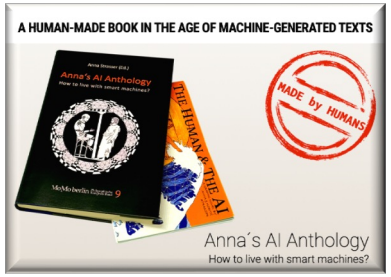
ANNA STRASSER (DenkWerkstatt Berlin, Germany)

DENKWERKSTATT BERLIN

INTERNATIONAL WORKSHOP

DIGIDAN

KÜNSTLICHE INTELLIGENZ Die große Verheißung



BOOK PROJECT
INBETWEENISM.
WHY EXISTING ETHICAL POSITIONS ARE OUTDATED WHEN FACING GENERATIVE AI TECHNOLOGY

Anna Strasser
https://www.denkwerkstatt.berlin

DENKWERKSTATT BERLIN

AI guidelines

No one wants to review and publish purely machine-generated text.

BUT

How can we handle the indistinguishability between human-created and machine-generated text?

➢ No detection algorithm / no expert / no naïve person can distinguish with certainty!

DETECTION SOFTWARE

some sources claim up to 98% accuracy → BUT suspiciously like advertising

- do not refer to experimental studies (Compilatio, 2024; Crossplag, 2024; Winston AI, 2024; Zero GPT, 2024)

scientific study by Weber-Wulff et al. (2023):

- 12 publicly available tools & two commercial systems (Turnitin, PlagiarismCheck): none was accurate or reliable
  - all scored below 80% accuracy, and only 5 over 70%
  - (consistent with N. Anderson et al., 2023; Demers, 2023; Elkhatat et al., 2023; Gewirtz, 2023; Krishna et al., 2023; Pegoraro et al., 2023; van Oijen, 2023; J. Wang et al., 2023)

➢ **How can a publisher verify the human authorship?**

Studies evaluating humans' limitations regarding indistinguishability

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

Brown et al. (2020). *Language Models are Few-Shot Learners*.

Gao et al. (2022). *Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers*

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a large language model of a philosopher.

Acceptable use cases

THE USE OF GENERATIVE TECHNOLOGY IS UNAVOIDABLE, WHICH USE CASES ARE ACCEPTABLE?

- Using DeepL, Grammarly, or ChatGPT
  - to improve drafts concerning language and style
    - to enhance the readability of a text
    - to suggest alternative phrasings

- We might even have no problem imagining that such machines are used as a muse to gain inspiration for future work.
  - Thinking tools for inspiration, improving language and style?

DigiDan could already be used *by me* as a generator of possible ways of putting things that I could then edit and revise, confident that I would not be wasting my time on irrelevant babble because of its competence. The composer/computer programmer David Cope created his EMI (Experiments in Musical Intelligence) as a tool to help him compose music when his muse was napping, and now, with a little more fiddling and improving, Anna Strasser may provide me with a highly efficient wordmonger tool that will suggest ways to me of putting my own convictions better! I can say at this point that this is not yet – in 2022 – a tool I can use, and I haven't used it (so far) in writing this book. But stay tuned.*

Dennett 2023, p. 276

Limitations of reliability

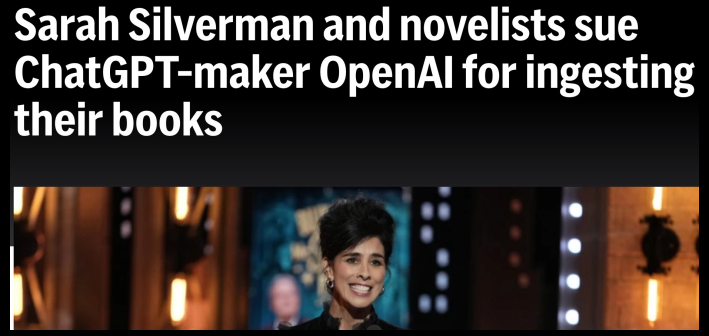
But how can we prevent the users from being victims of the unreliability of LLMs?

What about proofing that technology users are aware of the limitations regarding the reliability of generative AI technology?

Being well-informed (understanding & acknowledging) about these limitations of genAI technology like LLMs is crucial for responsible use.

Further critical issues:

- How can we ensure that the use of this technology does not contribute to new forms of plagiarism by using fine-tuned models of certain authors?
- And what about implicit violations of copyright issues?
- Can publisher avoid that the works of their authors are used as training data?



AI Guideline

Questions & Suggestions

- How can the publisher verify the human authorship?
  - Signature? / ORCID? ...
- Should the publisher require proof of being well-informed about the limitations regarding the reliability of generative AI technology?
  - e.g. the necessity to go through intense fact-checking
- Should authors be required to mention when they are using fine-tuned models on other authors and promise to make sure that they are avoiding plagiarism by referring to the original author?
- The publisher should inform about potential Copyright violations through LLM companies.
  - Clarify whether authors are okay to become training data.
- interesting issue
  - investigate whether OpenAI has used DeGruyter texts to train its LLMs
    - check WebText, WebText2 and Common Crawl
    - ask sample queries to the models and analyze the generated answers

References

Anderson, N., Belavy, D. L., Perle, S. M., Hendricks, S., Hespanhol, L., Verhagen, E., & Memon, A. R. (2023). AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. *BMJ Open Sport & Exercise Medicine*, *9*(1), e001568. https://doi.org/10.1136/bmjsem-2023-001568

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

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. https://doi.org/10.48550/arXiv.2107.00061

Compilatio. (2024, January 19). *What is the best AI detector?* Compilatio. https://blog.compilatio.net/en/blog/best-ai-detectors

Crossplag. (2024). AI Content Detector. *Crossplag*. https://crossplag.com/ai-content-detector

Dennett, D. (2023). *I've been thinking*. W. W. Norton & Company.

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

Demers, T. (2023, April 25). *16 of the best AI and ChatGPT content detectors compared*. Search Engine Land. https://searchengineland.com/ai-chatgpt-content-detectors-395957

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). *Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers* (p. 2022.12.23.521610). https://doi.org/10.1101/2022.12.23.521610

Gewirtz, D. (2023, October 19). *Can AI detectors save us from ChatGPT? I tried 5 online tools to find out*. ZDNET. https://www.zdnet.com/article/can-ai-detectors-save-us-from-chatgpt-i-tried-5-online-tools-to-find-out

Grynbaum, M. M., & Mac, R. (2023, December 27). The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. *The New York Times*. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). *Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense* (arXiv:2303.13408). arXiv. https://doi.org/10.48550/arXiv.2303.13408

Pegoraro, A., Kumari, K., Fereidooni, H., & Sadeghi, A.-R. (2023). *To ChatGPT, or not to ChatGPT: That is the question!* (arXiv:2304.01487). arXiv. https://doi.org/10.48550/arXiv.2304.01487

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a large language model of a philosopher. Mind & Language, 1–23. https://doi.org/10.1111/mila.12466

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. (2006). Kognition künstlicher Systeme. In Kognition künstlicher Systeme. De Gruyter. https://doi.org/10.1515/9783110321104

Strasser, A. (Ed.). (2024). Anna's AI Anthology. How to live with smart machines? xenomoi Verlag.

Strasser, A. (forthcoming). Inbetweenism. Why existing ethical positions are outdated when facing generative AI technology. de Gruyter.

Strasser, A., Sohst, W., Stepec, K., & Stapelfeldt, R. (Eds.). (2021). Künstliche Intelligenz – Die große Verheißung. (Vol. 8). xenomoi Verlag.

van Oijen, V. (2023, March 31). *AI-generated text detectors: Do they work? | SURF Communities*. https://communities.surf.nl/en/ai-in-education/article/ai-generated-text-detectors-do-they-work