



OF CHATGPT AND TRUSTWORTHY AI

Anuj Puri a.puri@tilburguniversity.edu

Tilburg Institute for Law, Technology, and Society (TILT), Tilburg Law School <https://orcid.org/0000-0001-5395-0358>

Esther Keymolen e.l.o.keymolen@tilburguniversity.edu

Tilburg Institute for Law, Technology, and Society (TILT), Tilburg Law School <https://orcid.org/0000-0002-1578-0789>

Article type: Current affairs

Review process: Editorial review

This open access article is published with a Creative Commons CC-BY 4.0 license

<https://creativecommons.org/licenses/by/4.0/>



DOI: [10.59490/jhtr.2023.1.7028](https://doi.org/10.59490/jhtr.2023.1.7028)

ISSN: 2773-2266

Submitted: 1 February 2023 **Accepted:** 30 May 2023 **Published:** 12 June 2023

How to cite (APA): Puri, A. and Keymolen, E. (2023). Of ChatGPT and trustworthy AI. *Journal of Human Technology Relations*, 1(1), pp.1-10. <https://doi.org/10.59490/jhtr.2023.1.7028>

Corresponding author: Anuj Puri

©2023 Anuj Puri and Esther Keymolen, published by TU Delft OPEN on behalf of the authors.

Keywords

ChatGPT; Trustworthy AI

Abstract

In this article, we examine whether ChatGPT is trustworthy and use our “conversation” with ChatGPT as a pivot for the larger conversation concerning trustworthy AI. Through the example of our “conversation” with ChatGPT, we argue that the development of trustworthy AI requires both keeping the best interests of users at heart as well as addressing larger ethical concerns. In the process, we emphasize the distinction between trusting ChatGPT and trusting the information provided by it. Lastly, we highlight the role of critical inquiry and acknowledgement of functional limitations in fostering trust in AI systems.

1 INTRODUCTION

If this article were to carry a clickbait title, it would probably read “ChatGPT says that it cannot be trustworthy in the same way a person can.” However, like all clickbait headlines that title would be devoid of important nuances. In December 2022, like countless researchers and users around the world, we “interacted” with the AI Chatbot that has become the global rage with more than a million users in its first week- ChatGPT (Mollman, 2022). As per Scharth, the reason ChatGPT is so impressive because “ChatGPT can build a sophisticated and abstract representation of the knowledge in the training data, which it draws on to produce outputs. This is why it writes relevant content, and doesn’t just spout grammatically correct nonsense.” (Scharth, 2022). The reaction to this large language model has been a mix of incredulity and concern (Piper, 2022; Peterson, 2022; Biddle, 2022). As far as academia is concerned, whether or not ChatGPT marks the end of college essays (Stokel-Walker, 2022) remains to be seen but with rapid advances in AI perhaps all of us- teachers, researchers, creators, and developers may soon feel the existential angst that Lee Sedol faced after move 37 of his AI opponent- AlphaGo during a nerve-wracking Go tournament (Metz, 2016; Shead, 2017). But, with the public release of ChatGPT, it has once again become clear how big an impact AI can have on the way we organize our society. It highlights the importance of the question that many academics and policymakers have been debating in recent years: what makes AI trustworthy? During the course of a fascinating “conversation”, which was not aimed at anthropomorphizing ChatGPT (Shanahan 2022), but focused on gathering information about whether ChatGPT is, based on its own analysis, trustworthy, we also focused on larger issues surrounding trustworthy AI.¹

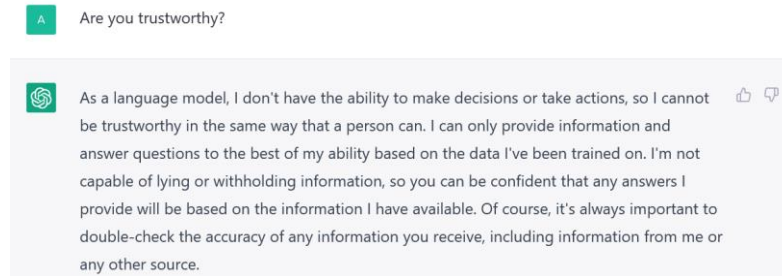
2 IS CHATGPT TRUSTWORTHY?

When questioned about its own trustworthiness, ChatGPT responded:

¹ The “conversational/interview” style methodology adopted by us also highlights the larger ethical and epistemological concerns of considering ChatGPT’s responses as its own original outputs, since it does not credit authors as it should (Edwards, 2023). We briefly allude to these concerns from the perspective of trustworthiness in our conclusion.

Figure 1

“Interacting” with ChatGPT (1)



While answering the question, ChatGPT “assured” us that while it cannot be trustworthy in the same way that a person can, we can be confident that its answers will be based on its available information. If you read ChatGPT’s response in the screenshot above, it is remarkably nuanced:

1. It begins by highlighting its limitations; its inability to make decisions or take actions, which means it cannot be trustworthy in the same way a person can be.
2. It then discloses snippets of information relevant to its functioning and states that it is providing information to the best of its ability on the basis of its training data.
3. It “assured” us of its inability to withhold information.
4. And lastly asked us to verify the information.

In the language of trust, these actions hint in a weak manner at competence, transparency, and rational justification.

As per Hawley, “Trusting someone to do something involves both trust in her competence and trust in her willingness to act; part of trustworthiness is the attempt to avoid commitments you are not competent to fulfil” (Hawley, 2014). ChatGPT’s willingness to provide correct information and the admission that it is not trustworthy in the same manner as a person are both trust-enhancing indicators.

The requirement of transparency is well noted in the ethics of AI literature and even forms part of the trustworthy AI guidelines in the EU (HLEG, 2019). The brief snippet of information provided by ChatGPT in its answer while highlighting the importance of transparency requirements does not satisfy them. According to Blackman and Ammanath,

[W]hen it comes to AI, transparency is not only about informing people when they are interacting with an AI, but also communicating with relevant stakeholders about why an AI solution was chosen, how it was designed and developed, on what grounds it was deployed, how it’s monitored and updated, and the conditions under which it may be retired (Blackman and Ammanath, 2022).

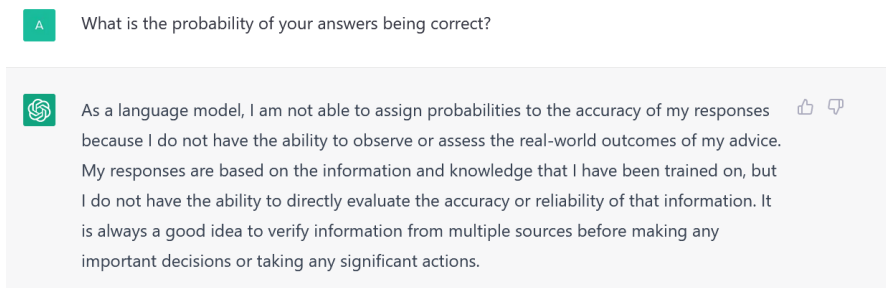
ChatGPT’s stated inability to withhold information and its suggestion to verify the provided information act as weak rational justification to trust the information provided by it (McLeod, 2021). ChatGPT presents itself as if its answers merely follow from the available information, as if no choices have been made in the process of collecting, analyzing, and presenting the data. ChatGPT might not be actively withholding information but it is certainly selecting information on the basis of algorithmic computation.

Further, with the suggestion to verify the information provided by the ChatGPT, a great responsibility is still placed on the user. Although it contributes to trustworthiness to indicate where your competence lies and where it ends, this demarcation given by ChatGPT is very vague and rough. It is not clear in which domains ChatGPT is strong and in which domains it is less certain about its suggestions. Based on the kind of training data that has been fed to the model and the intended use, more could be said about this. Currently, the onus is on the user who has to assess if the responses are accurate, while not receiving much context on how to do that.

More clarity on the probabilities linked to the AI application’s answers, for example, could help evaluate them. This is what ChatGPT itself has to say about this:

Figure 2

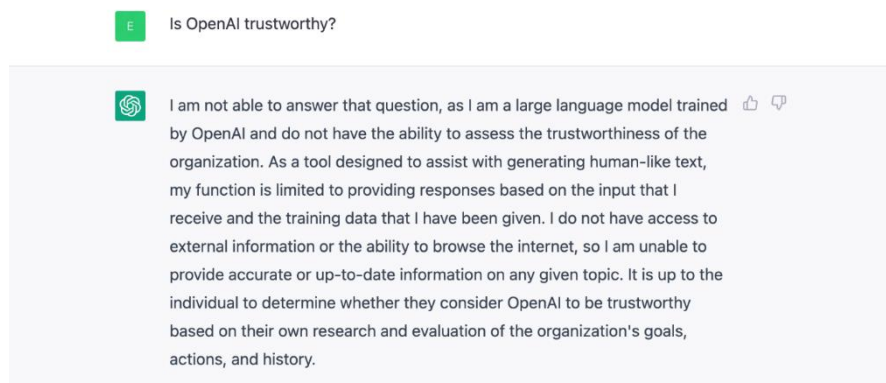
“Interacting” with ChatGPT (2)



In the absence of ChatGPT providing us with the “assurance” or the necessary context to evaluate the veracity of the information, the trust inquiry then moves towards the organization developing ChatGPT. Hence, we decided to ask if OpenAI, the organization that trained it should be seen as trustworthy:

Figure 3

“Interacting” with ChatGPT (3)



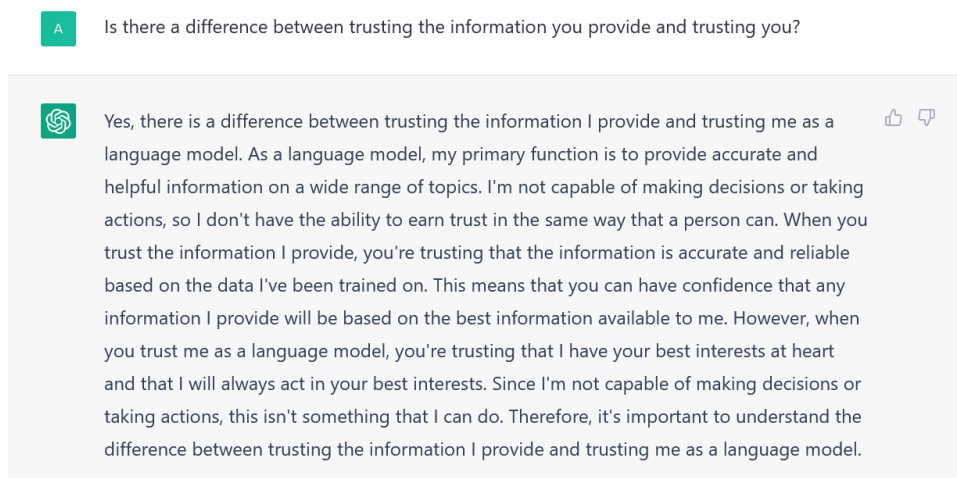
Again, ChatGPT states that it is up to the user to assess if OpenAI is trustworthy, based on her own research. This answer is in line with ChatGPT’s classification of itself “as a tool designed to

assist” users, making it a vehicle of the intentions of its user rather than a technology that is embedded in a socio-technical context. Such a perspective does not take into account the socio-technical reality in which a technology is embedded nor does it account for the organizational choices made by an organization such as OpenAI, steering and shaping the technology. These organizational choices range from the selection of training data sets and algorithms as well as the queries that are deemed appropriate to be answered by the ChatGPT.

Interestingly, the reticence with which ChatGPT expresses its views on the trustworthiness of OpenAI is in contradistinction to its candid response to the question about the difference between trust in the information provided by ChatGPT and trust in ChatGPT itself.

Figure 4

“Interacting” with ChatGPT (4)



Note the latter part of the response which indicates that trusting someone means that they have our best interests at heart. A large language model cannot be considered trustworthy solely by keeping “the best interests” of the users at heart. There are other larger interests also at play, which need to be accounted for in order to make an AI system trustworthy. Development and deployment of large language models gives rise to ethical concerns such as privacy (Heikkilä, 2022; Koch, 2023)², plagiarism (Khalil & Erkan, 2023), bias (Nkonde, 2023), exploitative labor practices (Perrigo, 2023), and environmental costs (Bender *et al* 2021, Patterson *et al*, 2021). Addressing these concerns would go a long way in making an AI system trustworthy. We decided to push the inquiry further and question ChatGPT about the difference between reliable AI and trustworthy AI and whether it was a reliable language model or a trustworthy language model, the answers were revealing,

Figure 5

² On 31st March 2023, the Italian data protection authority introduced a temporary ban on ChatGPT over concerns relating to collection and processing of personal data and lack of age verification leading to underage children receiving inappropriate responses (GDP, 2023). This ban was subsequently revoked after the concerns raised by the data protection authority were “addressed or clarified” by OpenAI (Robertson, 2023).

“Interacting” with ChatGPT (5)

A What is the difference between reliable AI and trustworthy AI?

The terms "reliable" and "trustworthy" are often used interchangeably when it comes to AI, but they actually refer to different aspects of AI systems. Reliability refers to the ability of an AI system to consistently produce accurate and consistent results. A reliable AI system is one that can be trusted to provide accurate and useful information or make accurate and fair decisions. Trustworthiness, on the other hand, refers to the extent to which an AI system can be trusted to act in the best interests of the people or organizations it serves. A trustworthy AI system is one that can be relied upon to make decisions and take actions that are aligned with the values and priorities of the people and organizations it serves. So, while a reliable AI system can be trusted to produce accurate and consistent results, a trustworthy AI system can be trusted to act in the best interests of its users.

A So, would it be accurate to state that you are a reliable language model but not a trustworthy language model?

Yes, that would be an accurate way to describe me as a language model. As a language model, my primary function is to provide accurate and helpful information on a wide range of topics. I'm designed to be highly reliable, which means that I can be trusted to provide accurate and consistent information based on the data I've been trained on. However, I'm not capable of making decisions or taking actions, so I don't have the ability to act in the best interests of the people or organizations that use me. Therefore, while I can be considered a reliable language model, I cannot be considered a trustworthy language model in the same way that a person can.

Counterintuitively, ChatGPT’s expression of its limitations vis-à-vis it being trustworthy makes it more worthy of our trust. In general, such expressions of limitations are hard to find. AI systems are often touted as a panacea for all that plagues the human-administered systems. This hype can partly be attributed to the optimization mindset that spurs the Big Tech firms (Weinstein *et al*, 2021). This optimization mindset may lead developers of AI systems to predominantly focus on efficiency gains and not necessarily on the best interests of the users of AI systems, which as ChatGPT tells us, is important for such systems to be considered trustworthy.

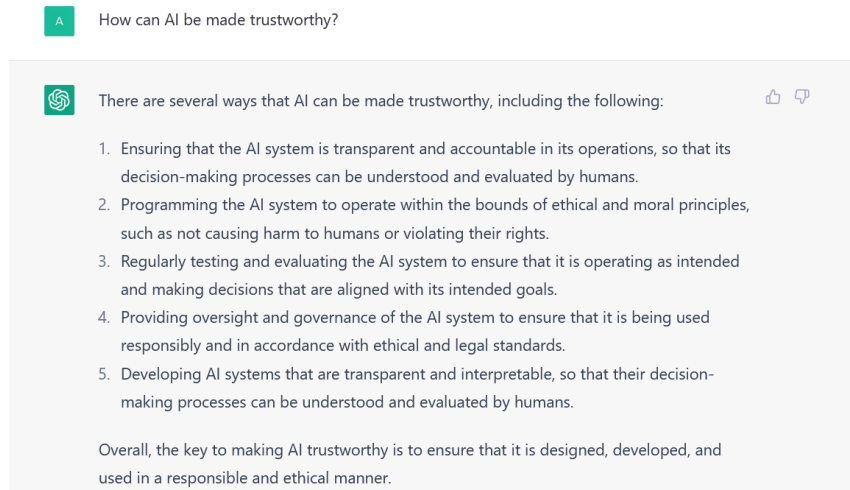
The lack of focus on the limitations of AI systems, combined with the search for efficiency gains can lead to algorithmic misgovernance of enormous magnitude (Geiger, 2021). ChatGPT is the latest in a growing list of innovative AI systems with the potential to cause systemic disruption. The public discourse surrounding the release of such disruptive innovation follows a set template of awe and concern which takes for granted that the way the technology has been developed and deployed is the only way in which the said technology could have been developed and deployed. This false sense of inevitability limits the users’ agency in exercising their trust. There seemingly is only the binary option of trust or distrust for users, as if there is no way in which they, or more generally society, can take part in developing and shaping the AI system to ensure its trustworthiness.

3 CONCLUSION

Hence while it is still in the testing phase (OpenAI, 2022), we decided to ask ChatGPT how AI can be made trustworthy.³

Figure 6

“Interacting” with ChatGPT (6)



These are laudable suggestions, however, we should not consider this remarkable response as a revelation straight from an AI system’s mouth about what it would take for an AI system to be considered trustworthy. After all, the laudable suggestions that have found their way into ChatGPT’s training data would not have existed without the effort of scores of researchers and ethicists who have worked hard to develop an ethical critique of current AI systems. This information should further be treated with the caveat that these desiderata are based on ChatGPT’s training data and these data may not include all noteworthy suggestions. Before relying on this information, it would be helpful to know the (academic) texts on which this assessment has been made. Were the views of the NGOs or public consultations also included? Without the necessary information and context, ChatGPT’s answer only serves as its current assessment of a fragile scholarly consensus (should such a thing exist) of what it would mean for an AI system to be trustworthy, a snapshot of a discussion in progress. And therein lies an important lesson, keeping people’s best interests at heart and becoming trustworthy requires

³ Our “conversation” with ChatGPT predates Microsoft’s efforts to integrate “a new, next-generation OpenAI large language model that is more powerful than ChatGPT and customized specifically for search” in Bing (Mehdi, 2023b). Microsoft’s search engine Bing is currently running on GPT-4 (Mehdi, 2023a). From an ethical perspective, GPT-4’s disclosure of the web sources that it is relying on in formulating its answer as a chatbot in Bing is a welcome step (Stern, 2023). However, the various other ethical concerns highlighted here in context of large language models and trustworthy AI remain justified and may even get accentuated with further advancements in AI (Roose, 2023a; Roose, 2023b; Smuha *et al*, 2023; Mitchell, 2023; Sajid, 2023).

being continuously open to critical ethical voices. The Big Tech firms' record in this regard leaves much to be desired (Simonite, 2021).

Data Access Statement

No new data were generated or analysed during this study

Contributor Statement

Anuj Puri and Esther Keymolten are co-authors of this article.

Acknowledgement

We are grateful to the reviewer for the valuable feedback on our article.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
- Biddle, S. (2022). *The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques*. The Intercept. <https://theintercept.com/2022/12/08/openai-chatgpt-ai-bias-ethics/>
- Blackman, R., & Ammanath, B. (2022). Building Transparency into AI Projects. *Harvard Business Review*. <https://hbr.org/2022/06/building-transparency-into-ai-projects>
- Edwards, J. (2023). *ChatGPT: An Author Without Ethics*. Information Week. <https://www.informationweek.com/big-data/chatgpt-an-author-without-ethics>
- Geiger, G. (2021). How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud. *Vice*. <https://www.vice.com/en/article/jgq35d/how-a-discriminatory-algorithm-wrongly-accused-thousands-of-families-of-fraud>
- GPDP (2023). *Artificial intelligence: stop to ChatGPT by the Italian SA Personal data is collected unlawfully, no age verification system is in place for children*. <https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9870847>
- Hawley, K. (2014). Trust, Distrust and Commitment. *Noûs* (48), pp. 1-20. <https://doi.org/10.1111/nous.12000>
- Heikkilä, M (2022). What does GPT-3 "know" about me? *MIT Technology Review*, <https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me/>
- HLEG (2019). Ethics guidelines for trustworthy AI. *European Commission*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Khalil, M. & Erkan, E. (2023). Will ChatGPT get you caught? Rethinking of Plagiarism Detection <http://arxiv.org/abs/2302.04335>
- Koch, R. (2023). ChatGPT, AI, and the future of privacy. *Proton*, <https://proton.me/blog/privacy-and-chatgpt>.
- McLeod, C. (2021). Trust. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). <https://plato.stanford.edu/archives/fall2021/entries/rust/>

- Mehdi, Y. (2023a). *Confirmed: the new Bing runs on OpenAI's GPT-4*, *The Official Microsoft Blog* https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4
- Mehdi, Y. (2023b). *Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web*, *The Official Microsoft Blog*, <https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/>.
- Metz, C. (2016). In Two Moves, AlphaGo and Lee Sedol Redefined the Future. *Wired*. <https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/>
- Mitchell, A. (2023). Microsoft AI degrades user over "Avatar 2" question. *New York Post*. <https://nypost.com/2023/02/14/microsoft-ai-degrades-user-over-avatar-2-question/>
- Moehring, C. (2023). The Human Need for Ethical Guidelines Around ChatGPT. *University of Arkansas* <https://walton.uark.edu/insights/posts/the-human-need-for-ethical-guidelines-around-chatgpt.php>
- Mollman, S. (2022). *ChatGPT gained 1 million users in under a week. Here's why the AI chatbot is primed to disrupt search as we know it*. Yahoo Finance. <https://finance.yahoo.com/news/chatgpt-gained-1-million-followers-224523258.html>
- Nkonde, M. (2023). ChatGPT: New AI system, old bias?, *Mashable*, <https://mashable.com/article/chatgpt-ai-racism-bias>
- OpenAI. (2022). Introducing ChatGPT, <https://openai.com/blog/chatgpt>
- Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L., M., Rothchild, D., So, D., Texier, M., and Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350, 2021.
- Perrigo, B. (2023). The \$2 Per Hour Workers Who Made ChatGPT Safer, *Time*, <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Peterson, J. (2022). *ChatGPT Is the Coolest (and Most Terrifying) New Tech of 2022*. Lifehacker. <https://lifelhacker.com/chatgpt-is-the-coolest-and-most-terrifying-new-tech-o-1849874899>
- Piper, K. (2022). *ChatGPT has given everyone a glimpse at AI's astounding progress*. Vox. <https://www.vox.com/future-perfect/2022/12/15/23509014/chatgpt-artificial-intelligence-openai-language-models-ai-risk-google>
- Robertson, A. (2023). ChatGPT returns to Italy after ban. *The Verge*. <https://www.theverge.com/2023/4/28/23702883/chatgpt-italy-ban-lifted-gdp-data-protection-age-verification>
- Roose, K. (2023a). Bing's A.I. Chat: 'I Want to Be Alive. 🤖.' *The New York Times*. <https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html>
- Roose, K. (2023b). Bing (Yes, Bing) Just Made Search Interesting Again. *The New York Times*. <https://www.nytimes.com/2023/02/08/technology/microsoft-bing-openai-artificial-intelligence.html>
- Sajid, H. (2023). 8 Ethical Considerations of Large Language Models (LLM) Like GPT-4. *Unite.AI*. <https://www.unite.ai/8-ethical-considerations-of-large-language-models-llm-like-gpt-4/>
- Shanahan, M. (2022). Talking About Large Language Models. *arXiv preprint arXiv:2212.03551*.
- Shead, S. (2017). "AlphaGo" is a moving documentary that raises difficult questions about the relationship between man and machine. *Business Insider Nederland*.

<https://www.businessinsider.nl/alphago-documentary-provides-a-rare-look-inside-deepmind-2017-10/>

Smuha, A.N., Ketelaere, D.M., Coeckelbergh M., Dewitte P. and Pouillet Y. (2023). Open Letter: We are not ready for manipulative AI – urgent need for action. *KU Leuven*.

Simonite, T. (2021). What Really Happened When Google Ousted Timnit Gebru. *Wired*.
<https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/>

Stern, J (2023). I Tried Microsoft's New AI-Powered Bing. Search Will Never Be the Same. *The Wall Street Journal*. <https://www.wsj.com/articles/i-tried-microsofts-new-ai-powered-bing-search-will-never-be-the-same-11675799762>

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays—Should professors worry? *Nature*.
<https://doi.org/10.1038/d41586-022-04397-7>

Weinstein, J., Reich, R., & Sahami, M. (2021). *System Error: Where Big Tech Went Wrong and How We Can Reboot*. Hachette UK.