Science in the era of ChatGPT, large language models and generative AI

Challenges for research ethics and how to respond

Evangelos Pournaras

1. Introduction

Since the release of popular large language models (LLMs) such as ChatGPT, the transformative impact of artificial intelligence (AI) on broader society has been unprecedented. This is particularly alarming for science and its conquest of truth (Chomsky/Roberts/Watumull 2023). Generative AI and, particularly, conversational AI based on language models set new ethical dilemmas for knowledge, epistemology and research practice. From authorship to misinformation, biases, fairness and safety of interactions with human subjects, research ethics boards need to adapt to this new era in order to protect research integrity and set high-quality ethical standards for research conduct (van Dis et al. 2023). This paper focuses on reviewing these challenges with the aim of laying foundations for a timely and effective response.

ChatGPT is an AI chatbot released in November 2022 by OpenAI. It is a Generative Pre-trained Transformer (GPT), a type of artificial deep neural network with a number of parameters in the order of billions. It is designed to process sequential input data, i.e. natural language, without labeling (self-supervised learning), but with remarkable capabilities for parallelization that significantly reduce training time. The model is further enhanced by a combination of supervised and reinforcement learning based on past conversations as well as human feedback to fine-tune the model and its responses (Stiennon et al. 2020; Gao/Schulman/Hilton 2022). Other corporations followed with similar chatbots such as the one of Bard by Google. Generative AI expands beyond text, for instance to, images, videos and code (Cao et al. 2023).

ChatGPT demonstrates powerful and versatile capabilities that are relevant for science and research. From writing and debugging software code to writing, translating and summarizing text, the quality of its output becomes indistinguishable from that of a human (Else 2023), while generating complex responses to prompts in a few seconds. Despite this success, AI language models suffer from hallucinations, an effect of producing plausible-sounding responses, which are nevertheless incorrect, inaccurate or even nonsensical. Illustratively, generative AI fails to abide by Asimov's three laws of robotics (Smith 2023): (i) Harmful outputs do occur (first law) (Wei/Haghtalab/Steinhardt; Davis 2023). (ii) Jailbroken prompts often result in both disobedience and harm (second law) (Wei/Haghtalab/Steinhardt 2023). (iii) New capabilities for autonomy, e.g., Auto-GPT (Yang/Hue/He 2023). Pervasiveness (integration on personal mobile devices) may create additional loopholes for conflicts to the first and second law (third law).

Disclaimers of ChatGPT state the following: "May occasionally generate incorrect information", "May occasionally produce harmful instructions or biased content", "Our goal is to get external feedback in order to improve our systems and make them safer", "While we have safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content. It is not intended to give advice", "Conversations may be reviewed by our AI trainers to improve our systems", "Please don't share any sensitive information in your conversations" and "Limited knowledge of the world and events after 2021".

Each of these disclaimers reveal alerting implications of using AI language models in science. They oppose core values to support research integrity such as the concordat (Universities UK 2020) of the UK Research Integrity Office (UKRIO): (i) *honesty in all aspects of research*, (ii) *rigor in line with disciplinary standards and norms*, (iii) *transparency and open communication*, (iv) *care and respect for all participants, subjects, users and beneficiaries of research* and (v) *accountability to create positive research environments and take action if standards fall short*.[1] Generative AI also challenges several of the Asilomar AI Principles (Future of Life Institute 2017).

Chomsky, Roberts and Watumull (2023) question the morality of asking amoral conversational AI moral questions, while Awad et al. (2018) show empirical evidence about the cross-cultural ethical variations and deep cultural traits

1 Cited from Universities UK 2020.

of social expectations from moral decisions of machines, i.e. the moral machine experiment. Generative AI runs the risks of copyright infringement and deskilling of early career researchers in scientific writing and research conduct (Gottlieb et al. 2023; Dwivedi et al. 2023). Security threats in online experimentation can 'pollute' human subject pools by replacing human subjects with conversational AI chatbots to claim compensations (Jansen/Jung/Salminen 2023; Wei et al. 2023). Without safeguards for such new sources of misinformation, data quality and research conduct can be degraded at scale.

AI language models also set foundational epistemological challenges addressing Karl Popper's seminal work on philosophy of science (Popper 2002 [1935]). Can AI language models assist us to make scientific statements that are falsifiable, or are they rather preventing us from doing so within their opaque nature? Are we addressing reality by relying our scientific inquiry on them, and which reality is this? Do over-optimized AI language models that are subject to Goodhart's law (Manheim/Garrabrant 2018) manifest irrefutable truth? And if so, do these models constitute the wrong view of science that betrays itself in its craving of being right?

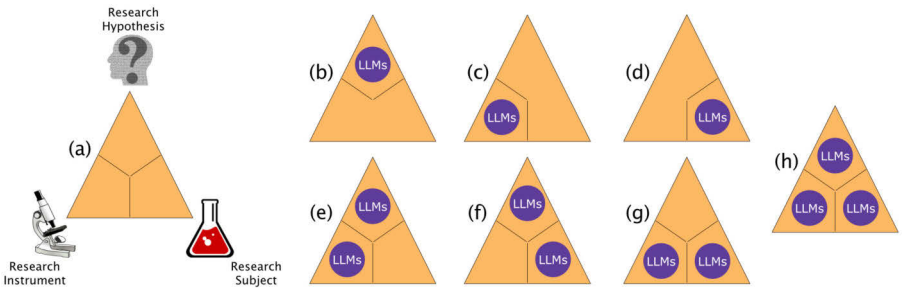
This paper dissects these questions with a focus on the research ethics review, although the discussion also finds relevance with regards to other facets of science such as education. To dissect the implications on science, the role of AI language models is distinguished as a *research instrument* and *research subject* when addressing a research hypothesis or question related or not to generative AI. Moreover, the ethical challenges of AI digital assistance to *scientists*, *human research subjects* and *reviewers* of research ethics are assessed. This scrutiny yields ten recommendations of actions to preserve and set new quality standards for research ethics and integrity as a response to the advent of generative AI.

This paper is organized as follows: section 2 reviews the different roles of generative AI in research design. Section 3 reviews the digital assistance provided by generative AI to scientists, participants and reviewers. Section 4 discusses emerging research ethics review practices in the era of generative AI. Section 5 introduces ten recommendations to respond to the challenges of research ethics review. Finally, section 6 concludes this paper and outlines future work.

2. The role of generative AI in research design

Within a research design serving a research hypothesis or question, generative AI can be involved as a research instrument or as a research subject, along with human subjects. This section distinguishes and discusses challenges and risks that may arise in these different contexts of a research ethics application. Figure 1 illustrates where generative AI such as large language models can emerge in a research design.

Figure 1: Generative AI such as large language models (LLMs) can be present in multiple stages of a research design within a research ethics application. Here, we depict all combinations: (a) No generative AI models are involved. (b) Generative AI models can be the motivation behind formulating a research hypothesis or question. (c) They can also be used as a research instrument to acquire knowledge. (d) They can also be the research subject itself, when interacting with human research subjects or when acting independently. (e)-(h) Generative AI models may be involved in multiple stages of the research design. In this case, it becomes imperative to distinguish their role at each phase to dissect research integrity and ethical dilemmas that may not be apparent anymore. Note that in (c), (d), and (g), where AI language models do not motivate a research hypothesis or question but they are involved as a research instrument or subject, research integrity and ethical risks are likely to arise. Image courtesy of the author.



2.1 Generative AI as a research instrument

ChatGPT is documented as an emerging research instrument capable of writing manuscripts for publication, often controversially featured as a coauthor

(O'Connor/ChatGPT 2022; ChatGPT Generative Pre-trained Transformer/Zhavoronkov 2022; Thorp 2023; Else 2023), writing software code (Dwivedi et al. 2023) and collecting data via queries (Dwivedi et al. 2023). Such tools are expected to come with capabilities for hypothesis generation in the future, including the design of experiments (van Dis et al. 2023; Dwivedi et al. 2023). Each of these instrumentations comes with different opportunities and challenges, including ethical ones.

During the design stage of research, including research ethics applications, there may be minimal support of AI language models on writing. However, the motivation of research, including literature review (Burger/Kanbach/Kraus forthcoming), generation of hypotheses, research questions as well as identifying ethical dilemmas, may be a result of interactions with conversational AI. Using the large capacity of conversational AI for knowledge summarization, these interactions can be systematized based on the Socratic method to foster intuition, creativity, imagination and potential novelty (Chang 2023).

However, often, creativity cannot be balanced with constraint (Chomsky/Roberts/Watumull 2023). At this stage, interactions with conversational AI require caution, running the risk of emulating or reinforcing a synergetic Dunning-Kruger effect (Gregorcic/Pendrill, 2023): conversational AI may rely on limited (or wrong) knowledge, which, while presented as plausible to humans with similar limited knowledge, may induce confirmation biases and diminish critical thinking. The mutual limitations of knowledge can be significantly underestimated in this context.

While research design choices may emerge from such interactions with conversational AI, a factual justification, a rigorous auditing process and moral judgments of these choices remain entirely under human premises (recommendation 1 and 8 in section 5). Finding reliable sources, revealing data sources, accurate contextualization of facts and moral framing are not attainable at this moment, as they require both cognitive capabilities, accountability and transparency that current AI language models lack (recommendation 1 in section 5). Whether existing ethics review processes are able to distinguish the risk level of research designs produced with the support of conversational AI as well as the mitigation actions, is an open question (recommendation 5 in section 5).

During research conduct, integrity and ethical dilemmas may arise when using the direct output of conversational AI (knowledge acquisition) to confirm or refute a hypothesis, especially when this hypothesis is not about the

AI system itself (see figure 1c, 1d, 1g and recommendation 4 in section 5). This output is in principle unreliable as it may contain incorrect or inaccurate information (Davis 2023). For instance, correct referencing may approach just 6 per cent (Blanco-Gonzalez et al. 2022). Moreover, AI language models tend to produce plausible content rather than content to be assessed as falsifiable, raising epistemological challenges (Popper 2002 [1935]). The reliability of AI language models as effective proxies for specific human populations is subject of ongoing research (Argyle et al. forthcoming).

Even if the output of AI language models is correct and accurate, it may not explain how such output is generated. For instance, there is often uncertainty to distinguish between lack of relevant data in the training set and failure to distill this data to credible information (van Dis et al. 2023). These models are usually black boxes with very low capacity to explain or interpret them. So far, this explainability is hard to assess for systems such as ChatGPT and Bard, which are closed and intransparent. This scenario may resemble an instrument collecting data exposed though to an unknown source of noise. Using instruments that have not passed quality assurance criteria may introduce various risks for users and work performed with such instruments and it is not different for AI language models. Standardized quality metrics are likely to arise for reporting to future research ethics applications (recommendation 6 in section 5), for instance, the 'algorithmic fidelity' that measures how well a language model can emulate response distributions from a wide spectrum of human groups (Argyle et al. forthcoming).

2.2 Generative AI as a research subject

The actual release of ChatGPT can be seen itself as a subject of research conducted by OpenAI with the aim to acquire user feedback that will improve AI language models. The initial interest lies in their actual capabilities to generate text and meaningful responses to user prompts. It also includes a discourse around their capabilities to perform calculations, write working code and jailbreaking via prompts that bypass the filters of its responses (Wei et al. 2023).

While these initial investigations are mainly experimental and anecdotal, a rise of empirical research on ChatGPT is ongoing (Dwivedi et al. 2023; Kim/Lee 2023; Bisbee et al. 2023), e.g., survey research. However, this outbreak of empirical research is to a certain extent a byproduct of releasing a closed AI black box with low capacity for explainability especially when the broader pub-

lic does not have access to the model itself or the exact data with which it is trained.

OpenAI and other corporations may benefit from such research as (free) crowd-sourcing feedback to calibrate their products, without sharing responsibility for doing so. Nonetheless, this may not be the original aims and intentions of scientists conducting such research. Such misalignment comes with ethical considerations on the value of this research and requires a critical stand by researchers and research ethics reviewers (recommendation 7 in section 5). While the methods of research on human subjects are well established (e.g., statistical methods, sociology, psychology, clinical research), the methods on AI subjects remain of different nature, pertinent to engineering and computer science. As human and AI subjects become more interactive, pervasive, integrated and indistinguishable, research ethics reviews need to account for (and expect) inter-disciplinary mixed-mode research methods (recommendation 2 in section 5).

3. Digital assistance by generative AI

AI language models can provide assistance to scientists, participants in human experimentation as well as to reviewers of research ethics applications. This section assesses ethical challenges pertinent to these beneficiaries.

3.1 AI-assisted scientist

As introduced in Section 2, the support of AI language models to scientists for literature review, writing papers, code, collecting data and performing experiments involves several challenges of integrity and ethics/moral. One question that may arise is how generative AI can contribute to the making of future scientists. Can they be part of the education of PhD students or will they result in deskilling, especially when students are not familiar with academic norms (Dwivedi et al. 2023)? Will such models be able to provide any level of self-supervision capability? The feasibility of research designs, success prediction of research proposals and reviewing manuscripts at early stages and before submission to journals, are some examples in which linguistics, epistemology and theory of knowledge set limits that for AI language models is hard to overcome (Chomsky/Roberts/Watumull 2023).

3.2 AI-assisted participant

Studying human research subjects assisted by AI language models requires a highly interdisciplinary perspective to dissect the ethical challenges and risks that may be involved (recommendation 2 in section 5). Such studies may aim to address the human subjects (i.e. social science), the AI language models when interacting with humans (i.e. computer science, decision-support systems), or both (e.g., human-machine intelligence). Design choices in AI systems for digital assistance to humans have direct ethical implications.

For instance, access to personal data for training AI models, centralized processing of large-scale sensitive information by untrustworthy parties and intransparent algorithms that reinforce biases, discrimination and informational filter bubbles pose significant risks. These include loss of personal freedoms and autonomy by manipulative algorithmic nudging, which participants may experience directly under research conduct, as well as broader implications in society (Hine 2021) related to environment, health and democracy (Pournaras et al. 2023; Asikis et al. 2021; Helbing et al. 2021; Helbing et al. 2023). The use of emerging open language models provides higher transparency to address some of these challenges (Patel/Ahmad 2023; Scao et al. 2022). Privacy-preserving interactions with AI language models, comparable to browsing with the DuckDuckGo search engine, are required (recommendation 3 in section 5).

Participants need to be informed about these risks when participating in such studies. For instance, information consent needs to account for any sensitive information shared during interactions with ChatGPT. Researchers do not have full control of the data collected in the background by OpenAI. As a result, participants need to be informed about the terms of use of AI language models. Moreover, responses by AI language models require moderation by researchers if they are likely to cause any harm to participants or special groups. Research ethics applications need to reflect and mitigate such cases (recommendation 9 in section 5).

3.3 AI-assisted reviewer

The support of generative AI to research ethics reviewers is a highly complex matter that perplexes both ethical matters within research communities as well as moral matters of individual reviewers. People do not share the same

judgments between the ethical choices of a human or a machine (Hidalgo et al. 2021).

AI language models show limited capabilities for ethical positioning, let alone moral positioning, possessing an apathy and indifference to implications of ethical choices (Chomsky/Roberts/Watumull 2023). They can endorse both ethical and unethical choices based on correct and incorrect information (ibid.). Nevertheless, they manage to influence users' moral judgments in a non-transparent way (Krügel/Ostermaier/Uhl 2023).

On the other hand, AI models can be used to effectively detect plagiarism or to perform pattern matching tasks that do not involve complex explanations or analysis of consequences. For instance, GPTZero is able to distinguish between text generated by humans vs. AI language models (Heumann/Kraschewski/Breitner 2023), which would be otherwise hard for reviewers to distinguish (Else 2023). Moreover, AI language models can assist reviewers, whose research background may be in a different discipline than the one of the proposed research. Summarizing necessary background knowledge and providing summaries in layman's terms can benefit research ethics reviewers (Hine 2021) as long as they remain critical on the generated output of AI language models.

As a result, AI language models are far from replacing reviewers in distilling ethical and moral implications of a research design, nevertheless, they can still play a role in the reviewing process by automating processes for pattern matching or making necessary background knowledge more accessible to reviewers, who may lack thereof.

4. Research ethics review practices

The need for regulatory and procedural reforms in research ethics review as a response to challenges of Big Data and data subjects dates back before generative AI (Ferretti et al. 2021; Hine 2021). Currently, the scope and practices of research ethics review are becoming broader and more multifaceted to cover the new alarming risks of generative AI. Two factors distinguish these research ethics review practices: (i) *scale of impact* and (ii) *stage of research*.

Institutional review boards for research ethics mainly address the impact of generative AI on human participants before the research conduct. Broader implications of the research on society are not explicitly addressed, although initial results from piloting an *Ethics and Society Review* (Bernstein et al. 2021) as a requirement to access funding show a positive impact (Bernstein et al.

2021). During research conduct, research ethics reviews mainly address any required adjustments in the research design rather than other unanticipated risks emerging from the application or new developments of AI.

Moreover, new research ethics review practices have recently been established for funding institutions (Bernstein et al. 2021), conferences and journals (Srikumar et al. 2022). These include (i) *impact statements*, (ii) *checklists* and (iii) *code of ethics or guidelines*. Impact statements include ethical aspects, questions and future positive or negative societal consequences, as well as identification of human groups, behavioral and socio-economic data. Checklists are used to flag papers for additional ethics reviews by an appointed committee, while code of ethics and guidelines support reviewers to flag papers that violate them.

While there is evidence that such practices can support panels to identify risks related to the harming of subgroups and low diversity (Bernstein et al. 2021), encouraging research communities to apply universal practices in different contexts and disciplines is a highly complex endeavor, given the current rapid AI developments and the unanticipated impact of these on society (recommendation 10 in section 5).

There are particular aspects of existing research ethics applications dealing with human aspects that are perplexed with the use of generative AI. These include individuals who can or cannot consent to terms of use and conditions of generative AI software, participants with disabilities, vulnerable groups and children, exclusion of certain groups, deception and incomplete disclosure, short and long term risks of participation, protection of personal data, anonymity and data storage. Research ethics review needs to address explicitly any additional risks involved in those aspects by using generative AI.

5. Ten recommendations for research ethics committees

This section introduces ten recommendations for research ethics committees. They distill the challenges and responses to AI language models involved in research ethics applications. They significantly expand on other earlier recommendations (Hine 2021) such as the one of World Association of Medical Editors (WAME) mainly addressing authorship, transparency and responsibility (Zielinski et al. 2023). They also constitute actions within the broader recommendations made for (i) studying community behavior and share learnings, (ii) expanding experimentation of ethical review and (iii) creating venues for

debate, alignment and collective action (Srikumar et al. 2022). The ten recommendations are summarized as follows:

1. Humans should always remain accountable for every scientific practice.
2. An interdisciplinary panel of reviewers should be employed to assess research ethics applications with elements on generative AI.
3. The use of generative AI models, their version, prompts and responses need to be documented and reported in any phase of the planned research. As a response, ethics reviews should detect potential inaccuracies, biases and inappropriate referencing. Mitigation by encouraging and promoting open generative models can improve accountability and transparency.
4. Research ethics applications that aim to address research hypotheses and questions out of the scope of generative AI, which do involve generative AI models as a research instrument or subject, are likely to involve research integrity and ethics issues and should be treated as high-risk applications.
5. Ethics review applications require new criteria and practices to distinguish low and high integrity risks in research designs produced with the support of generative AI. Determining appropriate mitigation actions to account for different risk levels is required.
6. Researchers who engage with generative AI in their research should report their countermeasures against inaccuracies, biases and plagiarism. Ethical review applications need to cover these risks.
7. The motivation and aim of research on generative AI should come with merit and go beyond testing of prompts lacking a rigorous scientific inquiry.
8. Auditing protocols are required for each input to generative AI models that are closed and proprietary, as a way to prevent sharing sensitive personal or proprietary information of researchers or participants.
9. Any output of generative AI that may harm participants or is sensitive to special groups requires moderation by researchers. Informed consent to relevant terms of use of generative AI models is required.
10. Communities on research ethics and regulatory bodies require to maintain an agreement on AI language models that can be used or should not be used in research. For instance, models that are obsolete, inaccurate, highly biased and violate values of science conduct shall be excluded, replaced or used with significant caution.

These recommendations should be used as an open and evolving agenda rather than a final list of actions. The current landscape of AI language models and research ethics remains multifaceted, rapidly changing and complex. Timely adjustments are needed as a response.

6. Conclusion and future work

To conclude, the challenges and risks of generative AI models for science conduct are highly multifaceted and complex. They are not yet fully understood, as developments are fast with significant impact and unknown implications.

Research ethics boards have a moral duty to follow these developments, co-design necessary safeguards and provide a research ethics review that minimizes ethical risks. A deep interdisciplinary understanding of the role that AI language models can play in all stages of research conduct is imperative. This can dissect ethical challenges involved in the digital assistance of scientists, research participants and reviewers.

The ten recommendations introduced in this paper set an agenda for a dialogue and actions for more responsible science in the era of AI.

Acknowledgements

Thanks to Maria Tsimpiri for inspiring discussions. Evangelos Pournaras is supported by a UKRI Future Leaders Fellowship (MR/W009560/1): *'Digitally Assisted Collective Governance of Smart City Commons – ARTIO'*, an Alan Turing Fellowship and the SNF NRP77 'Digital Transformation' project "Digital Democracy: Innovations in Decision-making Processes", #407740_-187249, the SNF NRP77 project 'Digital Transformation' project "Digital Democracy: Innovations in Decision-making Processes", #407740_187249 as well as the European Union, under the Grant Agreement GA101081953 attributed to the project H2OforAll – *Innovative Integrated Tools and Technologies to Protect and Treat Drinking Water from Disinfection Byproducts (DBPs)*. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. Funding for the work carried out by UK beneficiaries has been provided by UK Research and Innovation

(UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10043071].

List of references

Argyle, Lisa P./Busby, Ethan C./Fulda, Nancy/Gubler, Joshua R./Rytting, Christopher/Wingate, David (forthcoming): "Out of One, Many: Using Language Models to Simulate Human Samples." In: Political Analysis.

Asikis, Thomas/Klinglmayr, Johannes/Helbing, Dirk/Pournaras, Evangelos (2021): "How Value-sensitive Design Can Empower Sustainable Consumption." In: Royal Society open science 8/1, 201418.

Awad, Edmond/Dsouza, Sohan/Kim, Richard/Schulz, Jonathan/Henrich, Joseph/Shariff, Azim/Bonnefon, Jean-François/Rahwan, Iyad (2018): "The Moral Machine Experiment." In: Nature 563/7729, pp. 59–64.

Bernstein, Michael S./Levi, Margaret/Magnus, David/Rajala, Betsy A./Satz, Debra/Waeiss, Quinn (2021): "Ethics and Society Review: Ethics Reflection as a Precondition to Research Funding." In: Proceedings of the National Academy of Sciences 118/52 (https://doi.org/10.1073/pnas.2117261118).

Bisbee, James/Clinton, Joshua/Dorff, Cassy/Kenkel, Brenton/Larson, Jennifer (2023): Artificially Precise Extremism: How Internet-Trained LLMs Exaggerate Our Differences, SocArXiv Preprint (https://doi.org/10.31235/osf.io/5ecfa).

Blanco-Gonzalez, Alexandre/Cabezon, Alfonso/Seco-Gonzalez, Alejandro/Conde-Torres, Daniel/Antelo-Riveiro, Paula/Pineiro, Angel/Garcia-Fandino, Rebeca (2022): The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies, arXiv Preprint (https://doi.org/10.48550/arXiv.2212.08104).

Burger, Bastian/Kanbach, Dominik K./Kraus, Sascha (forthcoming): "The Role of Narcissism in Entrepreneurial Activity: A Systematic Literature Review." In: Journal of Enterprising Communities: People and Places in the Global Economy.

Cao, Yihan/Li, Siyu/Liu, Yixin/Yan, Zhiling/Dai, Yutong/Yu, Philip S./Sun, Lichao (2023): A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT, arXiv Preprint (https://doi.org/10.48550/arXiv.2303.04226).

Chang, Edward Y. (2023): "Prompting Large Language Models with the Socratic Method." In: 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Lay Vegas, NV, USA, pp. 0351–0360.

ChatGPT Generative Pre-trained Transformer/Zhavoronkov, Alex (2022): "Rapamycin in the Context of Pascal's Wager: Generative Pre-trained Transformer Perspective." In: Oncoscience 9, pp. 82–84.

Chomsky, Noam/Roberts, Ian/Watumull, Jeffrey (2023): "The False Promise of ChatGPT." In: The New York Times, March 8, 2023 (https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html).

Davis, Phil (2023): "Did ChatGPT Just Lie To Me?", January 13, 2023 (https://scholarlykitchen.sspnet.org/2023/01/13/did-chatgpt-just-lie-to-me/).

Dwivedi, Yogesh K./Kshetri, Nir/Hughes, Laurie/Slade, Emma Louise/Jeyaraj, Anand/Kar, Arpan Kumar/Baabdullah, Abdullah M./et al. (2023): "'So What If ChatGPT Wrote It?' Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy." In: International Journal of Information Management 71, 102642.

Else, Holly (2023): "Abstracts Written by ChatGPT Fool Scientists." In: Nature 613/7944, pp. 423–423.

Ferretti, Agata/Ienca, Marcello/Sheehan, Mark/Blasimme, Alessandro/Dove, Edward S./Farsides, Bobbie/Friesen, Phoebe/et al. (2021): "Ethics Review of Big Data Research: What should stay and what should be reformed?" In: BMC Medical Ethics 22/1, pp. 1–13.

Future of Life Institute (2017): "Asilomar AI Principles.", August 11, 2017 (https://futureoflife.org/open-letter/ai-principles/).

Gao, Leo/Schulman, John/Hilton, Jacob (2022): Scaling Laws for Reward Model Overoptimization, arXiv Preprint (https://doi.org/10.48550/arXiv.2210.10760).

Gottlieb, Michael/Kline, Jeffrey A./Schneider, Alexander J./Coates, Wendy C. (2023): "ChatGPT and Conversational Artificial Intelligence: Friend, Foe, or Future of Research?" In: The American Journal of Emergency Medicine 70, pp. 81–83.

Gregorcic, Bor/Pendrill, Ann-Marie (2023): "ChatGPT and the Frustrated Socrates." In: Physics Education 58/3, 035021.

Helbing, Dirk/Fanitabasi, Farzam/Giannotti, Fosca/Hänggli, Regula/Hausladen, Carina I./van den Hoven, Jeroen/Mahajan, Sachit/Pedreschi, Dino/Pournaras, Evangelos (2021): "Ethics of Smart Cities: Towards Value-sensitive Design and Co-evolving City Life." In: Sustainability 13/20, 11162.

Helbing, Dirk/Mahajan, Sachit/Hänggli Fricker, Regula/Musso, Andrea/Hausladen, Carina I./Carissimo, Cesare/Carpentras, Dino/et al. (2023): "Democracy by Design: Perspectives for Digitally Assisted, Participatory Upgrades of Society." In: Journal of Computational Science (https://dx.doi.org/10.2139/ssrn.4266038).

Heumann, Maximilian/Kraschewski, Tobias/Breitner, Michael H. (2023): ChatGPT and GPTZero in Research and Social Media: A Sentiment- and Topic-based Analysis, SSRN Preprint (https://dx.doi.org/10.2139/ssrn.4467646).

Hidalgo, César A./Orghian, Diana/Canals, Jordi Albo/De Almeida, Filipa/Martin, Natalia (2021): How Humans Judge Machines, Cambridge, MA: The MIT Press.

Hine, Christine (2021): "Evaluating the Prospects for University-based Ethical Governance in Artificial Intelligence and Data-driven Innovation." In: Research Ethics 17/4, pp. 464–479.

Jansen, Bernard J./Jung, Song-gyo/Salminen, Joni (2023): "Employing large language models in survey research." In: Natural Language Processing Journal 4, 100020.

Kim, Junsol/Lee, Byungkyu (2023): AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys, arXiv Preprint (https://doi.org/10.48550/arXiv.2305.09620).

Krügel, Sebastian/Ostermaier, Andreas/Uhl, Matthias (2023): "ChatGPT's inconsistent moral advice influences users' judgment." In: Scientific Reports 13/1, 4569.

Manheim, David/Garrabrant, Scott (2018): Categorizing Variants of Goodhart's Law, arXiv Preprint (https://doi.org/10.48550/arXiv.1803.04585).

O'Connor, Siobhan/ChatGPT (2022): "Open Artificial Intelligence Platforms in Nursing Education: Tools for Academic Progress or Abuse?" In: Nurse Education in Practice 66, 103537.

Patel, Dylan/Ahmad, Afzal (2023): "Google 'We Have No Moat, And Neither Does OpenAI'. Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI.", May 4, 2023 (https://www.semianalysis.com/p/google-we-have-no-moat-and-neither).

Popper, Karl R. (2002 [1935]): The Logic of Scientific Discovery, London and New York: Routledge.

Pournaras, Evangelos/Ballandies, Mark Christopher/Bennati, Stefano/Chen, Chien-Fei (2023): Collective Privacy Recovery: Data-sharing Coordination

via Decentralized Artificial Intelligence, arXiv Preprint (https://doi.org/10.48550/arXiv.2301.05995).

Scao, Teven Le/Fan, Angela/Akiki, Christopher/Pavlick, Ellie/Ilić, Suzana/Hesslow, Daniel/Castagné, Roman/et al. (2022): Bloom: A 176b-Parameter Open-access Multilingual Language Model, arXiv Preprint (https://doi.org/10.48550/arXiv.2211.05100).

Smith, Andrew (2023): "Asimov's Laws in Today's AI. ChatGPT and Other Generative AIs Graded.", June 19, 2023 (https://goatfury.substack.com/p/asimovs-laws-in-todays-ai).

Srikumar, Madhulika/Finlay, Rebecca/Abuhamad, Grace/Ashurst,Carolyn/Campbell, Rosie/Campbell-Ratcliffe, Emily/Hongo, Hudson/et al (2022): "Advancing Ethics Review Practices in AI Research." In: Nature Machine Intelligence 4/12, pp. 1061–1064.

Stiennon, Nisan/Ouyang, Long/Wu, Jeffrey/Ziegler, Daniel/Lowe, Ryan/Voss, Chelsea/Radford, Alec/Amodei, Dario/Christiano, Paul F. (2020): "Learning to Summarize with Human Feedback." In: Advances in Neural Information Processing Systems 33, pp. 3008–3021.

Thorp, H. Holden (2023): "ChatGPT is Fun, But Not an Author." In: Science 379, p. 313.

Universities UK (2019): The Concordat to Support Research Integrity, London: Universities UK (https://www.universitiesuk.ac.uk/sites/default/files/field/downloads/2021-08/Updated%20FINAL-the-concordat-to-support-research-integrity.pdf).

van Dis, Eva A. M./Bollen, Johan/Zuidema, Willem/van Rooij, Robert/Bockting, Claudi L. (2023): "ChatGPT: Five Priorities for Research." In: Nature 614/7947, pp. 224–226.

Wei, Alexander/Haghtalab, Nika/Steinhardt, Jacob (2023): Jailbroken: How Does LLM Safety Training Fail?, arXiv Preprint (https://doi.org/10.48550/arXiv.2307.02483).

Yang, Hui/Yue, Sifu/He, Yunzhong (2023): Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions, arXiv Preprint (https://doi.org/10.48550/arXiv.2306.02224).

Zielinski, Chris/Winker, Margaret/Aggarwal, Rakesh/Ferris, Lorraine/Heinemann, Markus/Lapeña, Jose Florencio/Pai, Sanjay/et al. (2023): "Chatbots, ChatGPT, and Scholarly Manuscripts – WAME Recommendations on ChatGPT and Chatbots in Relation to Scholarly Publications." In: Afro-Egyptian Journal of Infectious and Endemic Diseases 13/1, pp. 75–79.