Digital Society (2023) 2:48
https://doi.org/10.1007/s44206-023-00079-7

ORIGINAL PAPER

Norms for Academic Writing in the Era of Advanced Artificial Intelligence

Simon Friederich[1] · **Jonathan Symons**[2]

Received: 28 April 2023 / Accepted: 17 October 2023 / Published online: 2 November 2023

Abstract
If and when artificial intelligence systems become superhuman in more aspects of analytic reasoning, this will inevitably have a strong impact on the social organisation of science, including academic writing, reviewing, and publishing. We consider how norms of academic publishing should be adjusted as this happens. To do so, we propose four intuitively plausible desiderata that norms of academic publishing should fulfil in the age of increasingly advanced artificial intelligence (AI) and argue that there are no "quick fixes" to current norms that fulfil these desiderata. To indicate the scale of change needed to prepare academic publishing for the era of increasingly advanced AI, we tentatively sketch a more promising novel system of norms. Our proposal centres around the idea that AI systems should "sign off'' on statements that outline the human and AI contributions to academic research. We discuss possible challenges for this proposal and highlight the type of technological and regulatory infrastructure that would be needed to enable it.

Keywords Artificial general intelligence · Academic writing · Norms · Harm · Progress

1 Introduction

Developments in artificial intelligence (AI) generate challenges and opportunities for academic assessment, writing, reviewing, and publishing. Already today, large language models (LLMs) such as ChatGPT can produce texts on arbitrary topics with excellent linguistic expression. Increasing effort is being invested to

✉ Simon Friederich
s.m.friederich@rug.nl

Jonathan Symons
jonathan.symons@mq.edu.au

[1] University College Groningen, University of Groningen, Hoendiepskade 23/24, 9718 BG Groningen, The Netherlands

[2] Macquarie School of Social Sciences, Macquarie University, 25B Wally's Walk, Macquarie Park, NSW 2109, Australia

make AI systems ever more generally intelligent. Indeed, many AI experts expect that "artificial general intelligence" (AGI), which is superhuman in many, perhaps all, relevant aspects of cognition, will be developed within the next few decades (Avin, 2019; Grace et al., 2018). Achieving this is the explicit goal of OpenAI, the company that produces the ChatGPT model, which brought unprecedented attention to AI in 2022. To the extent that further efforts towards this goal will at least partly succeed, future systems will have much greater analytical rigour and originality than, say, ChatGPT. Public discussion is currently focused on changes to assessment that will be needed as ChatGPT creates new opportunities for students to cheat. However, in the coming decades, AI may transform academic research and publishing as well (Gendron et al., 2022); at least one publisher has already amended its rules to require that LLM use be disclosed and documented (Nature, 2023). But will such minimal changes be sufficient? The near-term possibility of AI with human, perhaps superhuman, capacities in more and more aspects of analytical reasoning calls for sustained reflection on how AI systems should be integrated in future academic writing and publishing practices.

Here, we consider how norms of academic review and publishing might be adjusted in the age of more advanced AI systems, beyond today's LLMs. Plausibly, such norms should minimise risks from the flaws, limitations, and potential for misuse while also capturing the individual and collective benefits from AI. Our discussion is conditional on the two-part assumption that (i) future AI systems will come to exceed human capacities in more and more key aspects of analytic reasoning and (ii) the precise future capabilities of AI are highly uncertain (Floridi, 2019). We are not in a position to judge how promising efforts to build "AGI" systems really are. *If* they succeed in full, this technological revolution will likely transform society—including education and academia—in ways that currently defy prediction. In today's human-centred academia, one part of the *raison d'être* for academic publishing is that it enables scientists to share their findings so that other people can use them (Hull, 1988). It is conceivable that knowledge production may eventually be taken over completely by AI systems. In that case, academic publishing might end or take an entirely different form, as its human-focused *raison d'être* would no longer connect to knowledge production but would be limited to publications' social functions. Our focus is both on the period *before* that happens (if it ever does) and on the various social purposes that academic publication, and the associated collective, multigenerational enterprise of scientific inquiry, achieves in addition to sharing research findings. In other words, our goal is to explore how norms of academic publishing should be adjusted during the period, whether it will be short or indefinite, in which ever more advanced AI are created but humans, and human social relations remain central within academia (Wilholt, 2013).

The paper is structured as follows: Section 2 collects four tentative desiderata for candidate novel norms of academic writing and publishing, based on trustworthiness, fairness, no-harm, and academic progress. Section 3 argues that the most obvious candidate adjustments to norms of academic publishing do not fare well against these desiderata. Finally, Section 4 sketches a novel system of public documentation of AI's contribution to academic publishing that shows more potential in satisfying our desiderata. The novel norms we propose require accompanying changes in

regulation and technological infrastructure. The point of proposing this system is not to endorse it as the best way to organise future academic publishing, but to indicate the scale of change needed to prepare academic writing and publishing for the era of advanced AI. The paper concludes in Section 5 with a short summary.

2 Some Desiderata for Norms of Academic Writing and Publishing

The goal for this section is to ground the following two sections, which argue that the arrival of advanced AI systems will require a major overhaul of academic writing and publishing norms. Here, we suggest a list of desiderata for how future norms of academic writing and publishing should address AI use. The listed desiderata are chosen to be intuitively appealing without further motivation. We do not make any attempt to derive them from first principles. They are loosely based on the ideas that norms of academic writing, publishing, reviewing, and editing should, insofar as possible, ensure the *trustworthiness* of what is published, contribute to *fairness* among human researchers, *avoid harm* to individuals and society, and enhance academic *progress*. These ideas align with traditional principles of scientific practice, which include transparency, honesty, originality, simplicity, and reproducibility (see Wilholt, 2013; ALLEA, 2023).[1] We do not claim that these desiderata are exhaustive; there may be good reasons for including additional desiderata, for instance, one related to *fostering epistemic diversity* (Heesen & Romeijn, 2019).

The desiderata we suggest are the following:

Adequate Attribution The norms should prevent people from obtaining credentials for achievements that are actually those of AI systems.

This desideratum is intended to contribute to fairness and avoid harm from cheating. It is also inspired by the value of progress both because adequate attribution helps increase chances that more skilled individuals obtain more opportunities and because education will likely need to preserve contexts where humans learn basic skills without the assistance of AI.

Enable Novelty/Disincentivise Redundancy The norms should make it more likely that novel and fruitful ideas, results, and arguments are generated, published, and given attention.

LLMs create a danger that a flood of linguistically polished but conceptually low-quality AI-enhanced papers might crowd out more substantive human-generated work, standing in the way of trustworthiness and hindering progress. This desideratum works against low barriers to the use of AI systems in academic writing. However, if advanced AI systems do become superhuman in key aspects of analytic reasoning, such systems could also decisively contribute to novel results. So this desideratum may also work in the other direction, against very high barriers to AI

[1] We thank an anonymous reviewer for this connection.

use in academic publishing. Together, these considerations (and the next desideratum) suggest that norms should be revised with a view to enabling social processes that will incentivise productive uses of AI and disincentivise non-productive uses.

Prevent Harm Inasmuch as possible, the norms should prevent harm.

Two types of harm have already become apparent from the use of AI. First, early AI applications designed to learn from existing datasets have tended to replicate, rather than challenge, existing patterns of discrimination (Chouldechova, 2017; Hasan et al., 2022). For example, in the Netherlands, an AI-assisted system to detect fraudulent reception of childcare benefits was at the heart of the "toeslagenaffaire". Tens of thousands of families were pushed into debt and their lives derailed, simply because they matched an AI-determined risk profile according to which, for instance, people with dual nationalities were more at risk of committing fraud (Heikkilä, 2022). Second, present-day LLM's can generate authoritative statements but lack a capacity for assessing their accuracy (Sobieszek & Price, 2022). This creates risks concerning unintentional generation of persuasive misinformation. In academic publishing, these two tendencies both point to the risk of discriminatory content being dispersed in the guise of authoritative academic writing.

Norms of academic publishing should preferably close any "responsibility gaps" when it comes to harm from AI use in academic publishing—especially in respect of *culpability* for harm and *active responsibility* to fulfil moral obligations in respect of AI system design (Matthias, 2004; Santoni de Sio & Mecacci, 2021). The producers and operators of AI technology in academic publishing must be incentivised to avoid harm that may come from such use.

Harms can also arise from differential access to AI systems. Ideally, publishing norms should not accentuate—and preferably should mitigate—inequities that arise from differential access to costly resources (notably, AI resources), though it is unclear to what extent publishing norms can address such inequities.

Finally, harm might arise from restricting academic freedom—understood as the freedom of individual researchers to teach and to learn, and of academic institutions to have a measure of autonomy (Altbach, 2001, 206). Human academic freedom could be restricted either through excessive AI influence over research trajectories or through unwarranted regulation of access to AI. For so long as human academic freedom continues to be valued, publishing norms must take account of each risk.

The final desideratum is not specifically related to the challenges posed by AI:

Good as Norms New norms of academic writing and publishing should be "good as norms".

By "good as norms" we mean having the potential to be adopted and internalised by the community. This requires several attributes: the norms must have sufficient correspondence with existing academic values that one can realistically expect the majority of scholars to (largely) endorse them; they must be sufficiently simple so that people are able to understand them and keep them in mind; and there should be

a realistic chance that norm violation will be detected and sanctioned. Norms that lack these features are impractical as they will not be adopted or will not attract high rates of compliance if formally adopted. These claims reflect experience from other contexts, where the "social fitness" of norms has been assessed in terms of fit with existing normative structure, fit with key actors' identities and the legitimacy of the actors promoting them (Bernstein, 2001, 184). Socialisation and internalisation of norms by individual actors is a complex social process which cannot be summarised by any single theory (Neumann, 2010). Evidence from social psychological research suggests that compliance with professional norms is highest where there are high levels of social consensus supporting them, where there are dense links and resource dependence within a community, and dense ties linking decision-makers to peers (Zelditch, 2001).

3 Four Quick Fixes and Their Problems

In this section, we consider four suggestions for "quick fixes" to the challenges posed by AI systems. We argue that such quick fixes will all prove inadequate if AI systems become superhuman in key aspects of analytic reasoning. The suggested "quick fixes" are as follows: first, accepting unrestricted AI use; second, complete prohibition of AI use; third, no change besides accepting AI systems as co-authors; and fourth, broadening the notion of plagiarism to include verbal output from AI systems.

3.1 First Suggestion: Accept Unrestricted AI Use

The first suggestion is to allow advanced AI systems, including today's LLMs, to be used in the same way as tools that check spelling or perform simple numerical calculations. Thus, their use in academic writing and publishing would be accepted without acknowledgement. In support of this suggestion, it might be argued that such tools can be of great help to non-native speakers of English in particular and so can play a role in mitigating inequalities. Indeed, there seems to be a consensus that unaided spelling skills and avoidance of cumbersome formulations are not the skills for which academic credentials should be awarded.

However, if applied to advanced AI and not simply to writing aids, this suggestion conflicts with all the desiderata we have listed. Already for today's LLMs, allowing unrestricted use in academic writing without acknowledgment of LLM use would be incompatible with "adequate attribution", it would be problematic for fairness in the assignment of academic credentials, it would open the doors to persuasive misinformation and prejudiced content, and it would incentivise redundancy by lowering the barriers to low-quality largely LLM-generated submissions. In addition, it might hamper scientific progress if credentials and, thereby, responsibilities are assigned to individuals who do not possess the analytic skills attributed to them. We expect these unjust impacts mean this quick fix would also not be able to garner widespread support.

3.2 Second Suggestion: Banning the Use of AI Systems Completely

It has been argued that LLMs are a form of "automated plagiarism" (Van Rooij, 2022) and that, hence, they should not be used at all in academic writing (and, a fortiori, publishing). A second potential "quick fix" is to generalise this reaction beyond LLMs and ban—or otherwise declare as unacceptable—the use of AI systems in academic writing and publishing. The *Science* journals have moved towards this proposal by declaring that "[t]ext generated from AI, machine learning, or similar algorithmic tools cannot be used in papers published in Science journals" (Science, 2023).

A difficulty with this suggestion is that it is unclear how it could be effectively implemented. If only exhortations were used to discourage the use of LLMs and other AI systems in academic writing, it seems doubtful that much would be achieved, making this suggestion fail by the standards of "good as norms". If, more aggressively, a ban was imposed on LLM use and measures were taken to actually implement it, for instance, by taking intrusive steps such as monitoring researchers' behaviour, the cure might easily end up creating more harm in academic freedom than it prevents. Moreover, if future AI systems do indeed come to exceed human capacities in more and more key aspects of analytic reasoning, a ban on their use would impede progress.

3.3 Third Suggestion: Without Further Change, Accept (or Require) That AI Systems Be Acknowledged as Authors

A third suggestion is to accept AI systems as (co-)authors, without further indicating how they contributed. Indeed, some academic papers already list ChatGPT as a co-author (Stokel-Walker, 2023).

Similar to the first suggestion of allowing unrestricted use of AI, this conflicts with all the desiderata we have listed. Allowing recognition of AIs as co-authors without requiring specification of AI's contribution would not involve "adequate attribution"; it would risk the crowding out effect of low-quality AI-assisted production and would do little to restrict the harms of persuasive misinformation and prejudiced content. Moreover, transgressions of the norms to recognise AI systems as co-authors would be difficult to detect. Most significantly though, by diluting the responsibility of human authors for academic output without requiring a statement of how AI systems have been utilised, this norm would fail to generate the kind of transparency that would foster the social practices of review and debate that will incentivise socially productive uses of AI. It is for this reason that one publisher has already banned recognising AIs as co-authors (Nature, 2023).

3.4 Fourth Suggestion: Including AI Output in the Definition of Plagiarism

A fourth suggestion, proposed by AI researcher Michael Black (Black, 2022), is to allow for AI-based input in academic publication, but only when properly attributed, for

instance, via some citation format similar to those currently used for input from human sources. This suggestion amounts to effectively expanding the definition of plagiarism such that it covers the unacknowledged output not only of any human but also any machine.

We believe that this approach is promising. Our own suggestion in Section 4 could be seen as a further development of it. However, as it stands, it raises the problem that transgression is extremely difficult to detect, perhaps impossible. Systems to detect AI use in writing based on statistical regularities have been developed, but, based on findings for current systems (Casal & Kessler, 2023; Gao et al., 2023), we suspect that at least for shorter passages, the signal of AI use will be too weak for conclusive detection. Moreover, the competitive development of more sophisticated AI and more sophisticated detection will likely have the inconclusive character of an arms race. AI use will become especially problematic if researchers get their key ideas and arguments from advanced AI systems and articulate them in their own terms. In this case the AI origin of ideas and arguments could no longer be detected based on linguistic criteria. This may create incentives for researchers to ignore a norm requiring acknowledgement of AI contributions. Moreover, at a point where AI will increasingly provide key ideas and arguments, treating those as inputs in the form of citations will not accurately reflect their importance. It may also be regarded as impractical because it enforces a line between human- and AI-based ("cited") content that could make the exposition of ideas and arguments cumbersome.

4 A Tentative Proposal: Mutually Approved Documentary Statement of AI Contribution

We outline a tentative proposal for new norms of academic publishing and accompanying regulatory measures that are more in line with the desiderata proposed in Section 2 than the "quick fixes" of Section 3. We then argue that this proposal illustrates why the scale of change needed will be greater than envisaged in current proposals (Black, 2022; Nature, 2023). The heart of our proposal is the idea of requiring a *documentary statement* (or section) that transparently outlines how a publication has been generated and what the human and AI contributions are. The documentary statement should be approved not just by the (human) authors but—once this becomes practicable—by the AI systems involved.

Documentary statements of a similar type are already included in many multi-authored publications today, specifying, for instance, who provided the research idea, who performed the data collection, who did the data analysis, etc. The guidelines recently announced by a major academic publisher seem to envision something along these lines (without, however, requiring AI approval or other mechanisms to promote compliance): they require that any LLM use be documented "in the methods or acknowledgments section" (Nature, 2023).

Requiring an AI-approved documentary statement goes a long way towards fulfilling the desideratum of *adequate attribution* to *support novelty* by minimising restrictions on the use of new technology and to *disincentivise redundancy* and *minimise harms* by facilitating scholarly review and analysis of more and less productive uses of AI systems. This scholarly analysis of productive uses of AI might be

conducted by other scholars after publication. However, we suspect that part of the gold standard of peer review might one day involve an adversarial AI system in the reviewing process, either distinct from the one used in creating the academic work or the same in adversarial mode (see Price & Flach, 2017).

To further deter the unattributed use of AI, we suggest that disclosure rules should apply to all publications. If an author writes without any AI support and wants this to be acknowledged, they must include an explicit statement to that effect. This would make it necessary to write a "lie of commission" if AI support actually was involved, which would both increase the psychological cost of norm violation and facilitate sanctioning if unattributed AI use is discovered (Levine et al., 2018). Since the norm is simple and consistent with emerging academic practice concerning acknowledgement of co-authoring, it satisfies some aspects of "good as norms". However, as outlined so far, it falls short of others in that it does not facilitate easy detection of transgression.

Resolving the problem of detection will likely not be possible unless new norms are supported by a regulatory response. To make it attractive and practical to conform to the norm of including a documentary statement, the process in which AI systems reach their output must be made transparent and trustworthy (Russo et al., 2023). Regulation including auditing and licensing procedures would be needed to ensure that licensed AI systems really are capable and reliable in "signing off on" how publications were actually generated. Preferably, the licensed systems would be "state-of-the-art" in their respective domains and superior to non-licensed publicly available systems, as this would further progress and disincentivise the unaccounted for use of non-licensed systems.

A system of licensing AI might also be helpful for preventing harm from misinformation, defamation, the entrenchment of biases, etc., as the licensing system could involve producer liability for harm attributable to AI contributions alongside other measures promoting a trustworthy AI ecosystem (Avin et al., 2021). Potentially though, at least part of the responsibility for harm arising from unacknowledged use of AI might remain with a publication's human authors in order to promote norm compliance.

Our proposal would not be without problems. The most principled concern, at least on one view, is that our proposal incrementally shifts power from humans to AI systems and/or the corporations developing them. This concern arises because our proposal requires that wherever AI is adopted as a tool in scientific research, these AI systems are also required to "have a say" describing how the research was generated. One may see this as a worrying step towards "AI takeover", which some regard as one of the most serious "existential risks" facing humanity (Bostrom, 2014; Russell, 2019). We sympathise with this worry. However, we note that much of this risk arises simply from developing AI capabilities so far that research which contains decisive contributions from AI may often be superior to research without AI contributions. The primary threat to humans' role in academia arises from the development of AI with superhuman capabilities, rather than from norms requiring verification of AI-assisted publications. To the extent that one finds worries about AI takeover plausible, these concerns would need to be addressed through regulatory efforts paralleling those applied to other technologies, e.g. medical

ones (Russo, 2023) and through the reflective choices of AI developers (Croeser & Eckersley, 2019). Such efforts may well end up limiting or at least slowing down the development of AI capabilities. However, this wider debate about AI regulation is beyond the scope of this paper.

Beyond such rather general concerns, a more concrete worry about the present proposal is that the infrastructure used to document the interaction between researchers and AI systems could potentially be used to monitor the researchers' activities more broadly and ultimately facilitate intrusions on academic freedom. Furthermore, regulations requiring licensing and transparency of AI use by academics may have unwelcome spillover impacts on other sectors' access to those AI systems and reinforce boundaries between those inside and outside academia. Such regulation may also slow, for better or worse, the pace of AI development and deployment within academic publishing. Those worried about AI takeover may in fact see this as a welcome aspect of our proposal.

Eliminating the harm arising from unequal access to AI systems is another challenge that cannot be resolved by academic norms alone. Instead, ensuring that access to AI systems that are suitable for research becomes as equitable as possible will likely be an increasingly important question of public policy and global social justice.

5 Conclusion

We have argued that if future advanced AI systems move further towards having broad human- (or superhuman-) level analytic skills, new norms and infrastructures for academic publishing will be needed. We outlined some tentative desiderata for such norms and sketched aspects of a candidate system of norms that might perform better than minor tweaks to the current norms—such as those that some publications have adopted in recent months. At the centre of our proposal is the idea that academic publications should include a *documentary statement*, certified by both human authors and—once this becomes possible—by the AI systems involved, about how the publication was generated.

When should such reforms be introduced? Current generations of AI do not seem to warrant this scale of change, since AI is not yet sufficiently powerful to make a significant human author-style contribution to academic research. However, three factors suggest that it is not premature for academic communities to begin deliberating on revised norms for the age of advanced AI. First, analogous changes (banning or requiring disclosure) of AI assistance are already being adopted in respect of student work, and it would be valuable to preserve consistency across all levels of academic integrity norms. Second, it may be easier for publishers to gain scholars' acceptance for a new disclosure norm at a time when compliance carries low costs (most papers will have nothing to disclose), rather than at a time when AI assistance is more widespread and compliance will be burdensome. Third, the speed of AI development suggests that such changes to the norms and regulations governing academic publishing may become urgently necessary within a few years or decades. The speed of social and policy change, by contrast, is relatively glacial. If norms and regulations are to keep pace with the changing capacities of AI, advance preparation will be needed.

Acknowledgements We would like to thank Benjamin Bewersdorf, Marian Counihan, Ryan Wittingslow, an anonymous referee, and the editor of *Digital Society* for helpful comments on earlier versions of this paper.

Author Contribution SF provided the research idea and the initial outline of the paper. SF and JS both drafted sections of the paper and revised multiple drafts. JS consulted ChatGPT in respect of several questions addressed in the paper, but since the outcomes were unsatisfactory, AI's contribution to the final text was limited to copy editing of spelling and grammar.

Funding No funding was used for this research.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Ethics Approval No research involving humans, animals, their data, or biological material was performed for this article. Therefore, no ethics approval was required.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

References

ALLEA. (2023). The European code of conduct for research integrity (Revised Edition 2023). Berlin. https://doi.org/10.26356/ECOC

Altbach, P. G. (2001). Academic freedom: International realities and challenges. *Higher Education, 41*, 205–219.

Avin, S. (2019). Exploring artificial intelligence futures. *Journal of AI Humanities, 2*, 171–193.

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., Anderljung, M., et al. (2021). Filling gaps in trustworthy development of AI. *Science, 374*(6573), 1327–1329.

Bernstein, S. F. (2001). *The compromise of liberal environmentalism*. Columbia University Press.

Black, M. (2022). Redefining plagiarism in the age of AI, *Perceiving Systems Blog*, 10 December, Redefining plagiarism in the age of AI | Perceiving Systems Blog (perceiving-systems.blog).

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics, 2*(3), 100068.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Croeser, S., & Eckersley, P. (2019). Theories of parenting and their application to artificial intelligence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 423–428.

Floridi, L. (2019). What the near future of artificial intelligence could be. *Philosophy and Technology, 32*(1), 1–15.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers". *NPJ Digital Medicine, 6*, 75.

Gendron, Y., Andrew, J., & Cooper, C. (2022). The perils of artificial intelligence in academic publishing. *Critical Perspectives on Accounting, 87*, 102411.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research, 62*, 729.

Hasan, A., Brown, S., Davidovic, J., Lange, B., & Regan, M. (2022). Algorithmic bias and risk assessments: Lessons from practice. *Digital Society, 1*(2), 14.

Heesen, R., & Romeijn, J. W. (2019). Epistemic diversity and editor decisions: A statistical Matthew effect. *Philosophers' Imprint, 19*(39), 1–20.

Heikkilä, M. (2022). AI: Decoded: A Dutch algorithm scandal serves a warning to Europe — The AI Act won't save us. *Politico*, March 30, AI: Decoded: A Dutch algorithm scandal serves a warning to Europe — The AI Act won't save us – POLITICO.

Hull, D. (1988). *Science as a process: An evolutionary account of the social and conceptual development of science*. Chicago University Press.

Levine, E., Hart, J., Moore, K., Rubin, E., Yadav, K., & Halpern, S. (2018). The surprising costs of silence: Asymmetric preferences for prosocial lies of commission and omission. *Journal of Personality and Social Psychology, 114*(1), 29.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183.

Nature. (2023). Editorial: Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature, 613*, 612.

Neumann, M. (2010). Norm internalisation in human and artificial intelligence. *Journal of Artificial Societies and Social Simulation, 13*(1), 12.

Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM, 60*(3), 70–79.

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Russo, F. (2023). What can AI learn from medicine? *Digital Society, 2*(2), 32.

Russo, F., Schliesser, E., & Wagemans, J. (2023). Connecting ethics and epistemology of AI. *AI and Society*. https://doi.org/10.1007/s00146-022-01617-6

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy of Technology, 34*, 1057–1084.

Science. (2023). *Science* Journals: Editorial Policies, Science Journals: Editorial Policies | Science | AAAS.

Sobieszek, A., & Price, T. (2022). Playing games with AIs: The limits of gpt-3 and similar large language models. *Minds and Machines, 32*(2), 341–364.

Stokel-Walker, C. (2023). ChatpGPT listed as an author on research papers: Many scientists disapprove. *Nature News*. https://doi.org/10.1038/d41586-023-00107-z

Van Rooij, I. (2022). *Against automated plagiarism*, personal blog post, 29 December, Against automated plagiarism – Iris van Rooij (irisvanrooijcogsci.com).

Wilholt, T. (2013). Epistemic trust in science. *The British Journal for the Philosophy of Science, 64*(2), 233–253.

Zelditch, M. (2001). Processes of legitimation: Recent developments and new directions. *Social PSychology Quarterly, 64*(1), 4–17.