Generative AI entails a credit-blame asymmetry

Sebastian Porsdam Mann,[1,*] Brian D. Earp,[2,*] Sven Nyholm,[3] John Danaher,[4] Nikolaj Møller,[2] Hilary Bowman-Smart,[5,6] Joshua Hatherley,[7] Julian Koplin,[6] Monika Plozza,[8] Daniel Rodger,[9,10] Peter V. Treit,[11] Gregory Renard,[12, 13, 14] John McMillan,[15] Julian Savulescu[16]

1. Bonavero Institute of Human Rights, Faculty of Law, University of Oxford
2. Uehiro Centre of Practical Ethics, University of Oxford
3. Faculty of Philosophy, Philosophy of Science, and Religious Studies, LMU Munich
4. School of Law, University of Galway
5. Ethox Centre, University of Oxford
6. Monash Bioethics Centre, Monash University
7. School of Philosophical, Historical, and International Studies, Monash University
8. Faculty of Law, University of Lucerne
9. Institute of Health and Social Care, School of Allied and Community Health, London South Bank University
10. Department of Psychological Sciences, Birkbeck, University of London
11. Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry
12. Applied AI Corporation, San Francisco
13. Frontier Development Lab, SETI, NASA
14. University of California, Berkeley
15. Department of Bioethics, University of Otago
16. Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore

* Corresponding authors.

Standfirst. *Generative AI programs can produce high-quality written and visual content that may be used for good or ill. We argue that a credit-blame asymmetry arises for assigning responsibility for these outputs and discuss urgent ethical and policy implications focused on large-scale language models.*

Word Count including standfirst: 2421

Introduction

The recent releases of large-scale language models (LLMs), including OpenAI's ChatGPT and GPT-4, Meta's LLaMA, and Google's Bard have garnered significant global attention, leading to calls for urgent community discussion of the ethical issues involved. LLMs generate text by representing and predicting statistical properties of language.[1] Optimized for statistical patterns and linguistic form rather than for truth or reliability, these models cannot assess the quality of the information they use.[2]

Recent work has highlighted ethical risks associated with LLMs, including biases arising from training data; environmental and socioeconomic impacts; privacy and confidentiality risks; the perpetuation of stereotypes, and the potential for deliberate or accidental misuse.[3,1] We focus on a distinct set of ethical

questions concerning moral responsibility – specifically blame and credit – for LLM-generated content. We argue that different responsibility standards apply to positive and negative uses (or outputs) of LLMs and offer preliminary recommendations. These include calls for updated guidance from policymakers that reflect this asymmetry in standards, transparency norms, technology goals, and establishment of interactive forums for participatory debate on LLMs.

Generative AI use statement

Note: a brainstorming document for this Comment was produced using ChatGPT. The authors then drafted their own manuscript incorporating some of the themes from the synthetically produced material. See Appendix A for details.

Negative versus positive responsibility

There are well-known challenges in assigning responsibility for actions taken by AI systems.[4] Many argue that a responsibility 'gap' arises from their use because of unclarity or indeterminacy with respect to who, if anyone, can be held responsible for their outputs. Others argue that existing legal and moral standards can trace responsibility back to designers and controllers. Much of this debate focuses on 'negative' responsibility gaps: who is to blame for negative outputs? What is interesting about LLMs is that they also raise concerns about 'positive' responsibility gaps: who, if anyone, can take credit for positive outputs?

Ascertaining whether an LLM's output is 'positive' or 'negative' may be difficult, and subject to reasonable disagreement. Many outputs will have positive consequences for some but negative consequences for others; their valence may also depend on the context, target, or aim of their use or application. To simplify, we use 'negative responsibility gaps' to refer to situations in which there is unclarity or indeterminacy about who, if anyone, deserves blame for LLM-generated content, while 'positive responsibility gaps' refer to analogous situations regarding the taking of credit.

Positive and negative responsibility gaps are asymmetrical

Traditional theories of blame, reflected in many legal standards, suggest that if we are reckless or negligent with respect to bringing about a negative outcome, even if we did not intend to do so, we can still be held responsible for it. In contrast, to deserve credit for a positive outcome, we must exert some effort, or display some form of talent, or make some sacrifice to bring it about.[5,6]

Applied to LLMs, this asymmetry suggests that society might be justified in holding persons accountable for deliberate or careless errors in generated text if they put such text to use in ways that negatively impact others, even if they did not put much skill and effort into generating that text. But we might not think people deserve credit for text generated without much skill and effort, such as ChatGPT-generated exam papers. This asymmetry implies that lessons from the existing literature on moral responsibility in AI may not always be applicable to LLMs. Instead, the discussion of positive responsibility gaps may be informed by the literature on the ethics of human enhancement, where the question is how technologically-enabled effort affects credit for performance in sports or elsewhere.[5,7]

Below, we discuss a range of positive and negative responsibility challenges arising from LLM use.

Negative outputs imply new responsibilities

The use of LLMs has significant potential to generate negative outcomes in a variety of sociotechnical contexts with challenging implications for the attribution of blame and responsibility. For instance, LLMs may facilitate the production of mis- and disinformation at scale by automating the generation of fake articles, Op-Eds, or position papers on controversial topics, further undermining trust in media and social institutions, and contributing to political cynicism and extremism.[8,9] LLMs could be used for 'astroturfing': automated generation of coherent, varied, and human-sounding snippets of text supporting particular objectives.[10] Readers of fake news stories generated by early versions of GPT found them almost as trustworthy as high-quality news stories.[11] Newer LLMs far exceed the capabilities of these earlier models and thus pose a greater and continually-evolving threat to civil society and political institutions.

This potential for LLMs to generate negative outcomes generates new responsibilities. In addition to holding individuals accountable for the accuracy of LLM-generated text they use, we recommend that users in institutional contexts be appropriately educated and trained in the responsible use of LLMs. Developing organisations should mitigate these risks, for example by emphasizing the potential for inaccurate outputs, placing disclaimers on generated text containing potentially misleading or erroneous information, and improving content moderation policies. Existing efforts to combat mis- and disinformation by increasing news literacy, disseminating fact-checking tools, and removing harmful generated content from social media, should be enhanced.

Journals should update authorship guidance

The International Committee of Medical Journal Editors (ICMJE) recommends four criteria for authorship. LLMs can arguably satisfy two of them, namely contributing to the design, analysis, or interpretation of data; and drafting the manuscript. The third criterion requires author approval of the version to be published and whilst some form of "approval" may be elicited from an LLM, this would not be in the intended human sense aimed at holding researchers accountable for the research and its possible implications. The remaining criterion requires an author to be accountable for all aspects of the work and for any questions regarding its accuracy or integrity. LLMs cannot currently meet this criterion. *Nature* and *Science* prohibit assigning authorship to LLMs for this reason.

Journal publishers and editors should therefore update authorship guidance to explicitly permit or exclude LLM authorship generally or in specific instances. Several articles already record ChatGPT as lead or co-author, despite the absence of any consensus guidance on authorship regarding non-human authors and contributors.

Transparency about the use of LLMs should be mandated

As mentioned, LLMs cannot assess the quality of information they use or produce. Those responsible for vetting text therefore need to pay extra attention to generated text, which they can only do when cognizant that such models have been used. LLM use is also relevant to allocating credit. If humans exerted little effort in producing a text, then the degree of credit and recognition they deserve for the resultant text is limited.

The minimum level of transparency for an article submission should include a statement of whether, how, and to what end an LLM was used. Human-generated prompts and resulting synthetic text should be submitted with the final submission as an appendix document. We included such a statement and

appendix when we submitted this manuscript and recommend that this should be considered standard practice. The standard we recommend for disclosure is the same as for humans. Where a human contributed substantially to a paper's design and conception, their role should be acknowledged regardless of whether they meet the criteria for authorship. Likewise an LLM that meets some but not all of the criteria for authorship should be acknowledged but not listed as an author. Using an LLM to generate a title or rephrase a small amount of text to make it more succinct would not warrant any kind of disclosure. We believe that transparency standards of this kind should be mandated by publishers and editors, as the *Nature* group of journals has recently done. Further guidance may be necessary to advise peer reviewers regarding use of LLMs in reviewing manuscripts under consideration.

Rights and interests depend on skill and effort

A related set of issues concern intellectual property (IP) rights, including copyright and patent protection, which are based on conceptions of invention, creativity, and intellectual creation and do not, traditionally, recognize nonhumans as potential rights holders.[12] Similarly, the distinct but related category of authors' rights – rights to recognition as being the author of a work; not having that work misrepresented in ways prejudicial to one's interests; and to make a decent living off one's work – are protected at the international level *inter alia* by international human rights law, which applies only to humans.

Whether humans generating output via LLMs can obtain IP protection depends on the extent to which they demonstrate labour or creativity.[13] The degree of human input is particularly interesting where model output has been adapted either via in-context learning or by fine-tuning. Assessing authorial claims requires transparency about the use of LLMs. Where human input does not reach a threshold of significance, the work remains authorless.[14] To resolve such delineation issues, new models such as 'contributorship' for generated works and labelling duties should be explored.[15]

Context matters for credit and responsibility

Issues of credit and responsibility for LLM outputs will vary by context according to the aims, purposes, and importance of specific areas of application. A prominent example is the use of LLMs in education. Positive uses are possible. LLMs could be used for didactic purposes, to help generate outlines, to critique essay plans, and to explain complicated topics. Given the error-prone nature of LLMs, potential didactic uses should be carefully considered and undertaken with the greatest possible amount of educational supervision. Their sometimes factually incorrect output could be used to teach critical thinking and the importance of factual verification. However, overreliance on such models may dull writing and analysis, prevent the acquisition of critical thinking skills, and result in erroneous belief formation.

One solution to this would be to change assessment style to take into account students' potential use of LLMs (e.g. increased use of oral exams or in-person exams, which may allow some use of such tools). Alternatively, more invasive monitoring of students' writing progress could be undertaken. However, attempts to monitor students to prevent cheating (e.g. 'online proctoring') raised significant concerns about intrusions on students' privacy as well as cybersecurity risks.[16] Alternatively, educational institutions may need to fundamentally re-evaluate their underlying pedagogical philosophies (e.g., shifting focus to the process of learning rather than the evaluation of outputs) and restructure courses around a range of possible alternative and non-traditional assessments.[17]

Educational institutions may also need to consider investing in software that is more reliably able to detect text produced by LLMs (see: Syme 2022; Wiggers 2022). Current technology remains error-prone and is not sufficient for standalone detection of synthetically produced text. Furthermore, attempts to detect LLM use may simply result in an ongoing and potentially futile arms race. Given these potential problems, educational institutions should amend their academic misconduct guidance to address the use and misuse of LLMs, or modify their examination procedures. Such amendments should also reflect the potentially positive educational uses of LLMs.

Credit-blame asymmetry may lead to achievement gaps

Since the Industrial Revolution, automating technologies have made workers redundant in many industries, particularly agriculture and manufacturing. The recent assumption has been that creatives and knowledge workers would remain much less impacted by these changes in the near-to-mid-term future. Advances in LLMs challenge this premise.

How these trends will impact human workforces is a key but unresolved question.[18] The spread of AI-based applications and tools such as LLMs will not necessarily replace human workers; it may simply shift human workers to tasks that complement the functions of the AI. This may decrease opportunities for human beings to distinguish themselves or excel in workplace settings. Their future tasks may involve supervising or maintaining LLMs that produce the sorts of outputs (e.g., text or recommendations) which skilled human beings were previously producing and for which they were receiving credit. Consequently, work in a world relying on LLMs might often involve "achievement gaps" for human beings: good, useful outcomes will be produced, but many of them will not be achievements for which human workers and professionals can claim credit.[7]

This may result in an odd state of affairs. If responsibility for positive and negative outcomes produced by LLMs is asymmetrical as we have suggested, humans may be justifiably held responsible for negative outcomes created, or allowed to happen, when they or their organizations make use of LLMs. At the same time, they may deserve less credit for AI-generated positive outcomes, since they may not be displaying the skills and talents needed to produce text, exerting judgment to make a recommendation, or generating other creative outputs.

Recommendations for adapting responsibility standards

As LLMs continue to advance, it is important that policymakers, regulators, and those responsible for standards and norms in affected areas update and implement guidance, rules and laws aimed at maximizing the benefits and minimizing the harms of these powerful tools. One way to do this is to revise and adapt responsibility standards to address the new technologies.

These efforts should reflect the positive potential of LLMs as well as their potential harms and be informed by existing normative instruments, consultative and deliberative processes, and relevant debates in adjacent areas, such as human enhancement and intellectual property rights.

LLM outputs should be vetted before use, labelled, and should not yet be applied in important decision-making contexts. Transparency about the degree to which individuals rely on LLMs in generating outputs should be mandated. Research into LLM and AI theory, watermarking, improving training data quality,

and enhancing explainability and truth-tracking should be encouraged, as it could help ameliorate harms and facilitate responsible beneficial uses.

As recognized in a recent *Nature Machine Intelligence* editorial, AI/LLM developers could take inspiration from the self-regulation of fields such as biomedicine. Although initial efforts are underway to highlight the importance of ethical regulation, for example in publishing and in establishing analogues to institutional review boards or research ethics committees in AI research,[19] much more internal work on self-regulation and transparency is needed to establish trust that these technologies are developed in an ethical, responsible manner.

In addition to self-regulation, LLM governance could benefit greatly from input by other fields and stakeholders, such as civil society and the public. We recommend that forums for discussion of the ethical, legal, and societal implications of LLMs be established and their use promoted. These could take various forms, such as conferences, workshops, community gatherings and online platforms, and could provide opportunities for researchers, policymakers, and other stakeholders to engage in dialogue and exchange ideas on the responsible development and use of these technologies. To ensure that these forums serve their democratic purpose, they should be inclusive and representative of a diverse range of viewpoints.

References

1. Weidinger, L. et al. 2021, *ACM FAccT* 214–229 (2021).

2. Shanahan, M. Preprint at: https://arxiv.org/abs/2212.03551 (2022).

3. Bender et al. *ACM FAccT* 610–623 (2021).

4. Nyholm, S. *Sci. Eng. Ethics* **24,** 1201-1219 (2018).

5. Maslen, H., Savulescu, J. & Hunt, C. *Australas. J. Philos.* **98**, 304-318 (2019).

6. Nyholm, S. (2023). *This is Technology Ethics: An Introduction*, ch. 6 (Hoboken: Wiley-Blackwell, 2023).

7. Danaher, J. & Nyholm, S. *AI Ethics* **1**, 227-237 (2021).

8. Rini, R. *Philos. Impr.* **20**, 1-16 (2020).

9. Lazer, D.M. et al. *Science* **359**, 1094-1096 (2018).

10. Radford, A. et al. Technical report, OpenAi, 2019.

11. Kreps, S. & McCain, M. *Foreign aff.* (2019).

12. Bosher et al. Report at: https://www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/call_for_comments/pdf/org_brunel.pdf (2020).

13. Eshraghian, J.K. *Nat. Mach. Intell.* **2**, 157-160 (2020).

14. Ginsburg, J.C. & Budiardjo L.A., *Berkeley Tech. L. J.* **34,** 343-448 (2019).

15. Miernicki, M. & Ng (Huang Ying), I., *AI Soc.* **36,** 319-329 (2021).

16. Coghlan, S. Miller, T. & Paterson, J. *Phil. & Tech.* **34**, 1581-1606 (2021).

17. Archambault, L. Leary, H. & Rice, K. *Educ. Psychol.* **57**, 178-191 (2022).

18. Korinek, A. In: Dubber, M.D., Pasquale, F. & Das, S. (eds.) *The Oxford Handbook of Ethics of AI,* 475-492 (Oxford Univ. Press, 2020).

19. Srikumar, M et al. *Nat Mach Intell* **4**, 1061–1064 (2022).

Acknowledgements: We wish to thank an anonymous reviewer for very helpful and timely suggestions for improvements to an earlier version of this article.

Competing interests: We declare no competing interests.