Holding Large Language Models to Account

Ryan Michael Miller
Philosophy Department
Université de Genève
Geneva, Switzerland
Ryan.Miller@unige.ch

***Abstract*— If Large Language Models can make real scientific contributions, then they can genuinely use language, be systematically wrong, and be held responsible for their errors. AI models which can make scientific contributions thereby meet the criteria for scientific authorship.**

Keywords— Large Language Models, authorship, responsibility, reference, hallucinations

I. The AI Authorship Controversy

Large Language Models (LLMs) are transformer-based deep-learning neural networks with hundreds of billions of parameters trained by self-supervised learning on large text corpora to perform next-token prediction. OpenAI's November 2022 public release of their 175-billion parameter GPT-3.5 model trained with Proximal Policy Optimization [1] made available for the first time an AI with human-level performance on a wide range of cognitive tasks [2] and its 4,096 token context window (~3000 words for prompt + response) allowed a wide domain of application [3]. The March 2013 release of GPT-4, with a maximum 32,768 token (~24,000 word) context window, performance at the upper end of the human scale on many cognitive tasks, and twice the measured factual reliability [2] has only increased the possible uses.

One such use of LLMs is the production of scientific research, with hundreds of papers appearing on preprint servers with AIs listed as co-authors, some of which have been published with that authorship credit after peer review [4]. Since use of LLMs not only speeds the writing [5] and revision [6] process but also helps with literature review[7], algorithm development, data analysis, hypothesis generation [8] and even creativity [9] and argumentation [10], we can expect such use to continue to grow. Unlike in the case of previous computerized text generators like SCIgen [11] which merely slipped gibberish through sham or slipshod refereeing processes [12]–[15], LLMs generate text which can be genuinely useful and is sometimes undetectable even by dedicated referees [16], [17]. Until recently, the vast majority of ethical concern around LLM authorship has been about plagiarism [6], [18], [19]. Consequently, accountability efforts have focused on ways to deter or detect LLM use in scientific writing [16], [17].

Giving the LLM authorship credit neatly sidesteps plagiarism issues, however: if the LLM is listed as an author of the paper, then there can be no allegation that the other authors plagiarized from the LLM or that the contribution of the LLM lacked transparency. This transparency is further increased for journals which use a structured author contribution statement [20], [21] or contributor roles taxonomy [22], which would list the exact research and writing contributions made by the LLM to the final published product. Current suggestions for making such roles more specific [23] only raise the likelihood that LLMs would qualify for authorship. Furthermore, while not all actual writers of scientific literature must receive authorship credit in all disciplines according to prevailing ethical standards [24], almost one third of publication ethics codes and more than half of Social Sciences Citation Index journals *require* authorship credit for all participants in drafting and revising the text [25]. Even in the remainder which also require scientific contributions, LLMs may qualify given the capabilities discussed above. Certainly, in many of the existing exemplars of published peer-reviewed scientific work with LLM authorship credits the LLM must make a "substantial scientific contribution" if the work has one at all, since the vast majority of the text and nearly all of the argument comes in the form of text from unedited LLM token output. Without crediting LLMs as authors it is difficult to see how papers where they contribute substantially could comply with the International Committee of Medical Journal Editors (ICMJE) original fourth principle for authorship [26]:

> *Each part of the content of an article critical to its main conclusions and each step in the work that led to its publication [(1) conception or design of the work represented by the article, or analysis and interpretation of the data, or both; (2) drafting the article or revising it for critically important content; and (3) final approval of the version to be published] must be attributable to at least one author.*

Cases where LLMs have received authorship credit have involved every one of these steps [27], [28].

Nonetheless, the influential Committee on Publication Ethics (COPE) and World Association of Medical Editors (WAME) have called for banning AI authorship on the grounds that AIs "cannot take responsibility" for their output [29], [30], and this call has been heeded by *Nature* [31] while other influential journals have banned AI authorship without giving explicit reasons [32], [33]. ChatGPT's authorship has been retracted in one case on this basis [34]. COPE's standard combines a general responsibility test with a long history in the publication ethics literature going back to [26] with a more recent legal personhood test supposedly required for "assert[ing] the presence or absence of conflicts of interest" and "manag[ing] copyright and license agreements" [29]. WAME spells out the latter, legal test as a matter of the corporate form chosen by OpenAI and its disclaimer of responsibility [30], which are obviously contingent matters not essential to AI. Indeed, various forms of legal personhood have already been proposed for algorithms which would allow them to enter into contracts [35]–[38] and corporations may soon be forced to assume liability for the AIs they create [39]–[41]. LLMs are as capable of asserting the presence or absence of conflicts of interest as they are of asserting anything else. Philosophical interest in COPE's new standard thus lies with its responsibility test, which is supposed to be an addition to (or even restriction of) the "substantial scientific contribution" standard for authorship which LLMs cannot meet even if or when they meet the latter standard.

COPE's responsibility standard goes back to ICMJE's original first principle for authorship [26]:

> *Each author should have participated sufficiently in the work represented by the article to take public responsibility for the content…[which] means that an author can defend the content of the article, including the data and other evidence and the conclusions based on them. Such ability can come only from having participated closely in the work represented by the article and in preparing the article for publication. This responsibility also requires that the author be willing to concede publicly errors of fact or interpretation discovered after publication of the article and to state the reasons for error. In the case of fraud or other kinds of deception attributable to one or more authors, the other authors must be willing to state publicly the nature and extent of deception and to account as far as possible for its occurrence.*

LLMs like ChatGPT manifestly both defend their output [9] and apologize for mistakes while giving reasons for their occurrence [6], [9] as well as identify particular human co-authors by their writing and offer criticisms [6]. WAME additionally references the current ICMJE standard that all authors must provide "Final approval of the version to be published" [42] as a reason that AIs cannot meet the general responsibility test [30]. While some publications with LLMs listed as co-authors may be suspect in this regard [27], ChatGPT's unwillingness to co-author is likely a result of its Reinforcement Learning from Human Feedback (RLHF) and is obviously not essential to LLMs. The COPE/WAME/*Nature* general responsibility test for authorship is thus best understood as a normative claim rather than a legal or behavioral one. ICMJE's "criteria are not intended for use as a means to disqualify colleagues from authorship who otherwise meet authorship criteria" [42], so the question is whether LLMs which meet the research and writing standards for authorship are able "to be accountable" in some normative sense. This is a fundamentally philosophical question.

The philosophical response to COPE's general responsibility test for AI authorship has been mixed. Wiese grants that current AIs are insufficiently agential to meet this constraint, but holds that future "strong artificial consciousnesses" which observe the Free Energy Principle would exhibit the relevant normative properties [43]. Jenkins and Lin, by contrast, argue that many uncontroversial human authors (e.g., deceased ones) also cannot take responsibility, so that only the research and writing standards are appropriate [44]. Another similar approach suggests that responsibility for scientific publications is best understood as irreducibly collective among the authors [45] so that AIs are accountable as part of a system with relevantly-situated humans [46], i.e. co-authors. On this approach, if there is a single human co-author to take responsibility, then the authorship team as a whole does, and further accountability is required of the AI. I take a third approach: *if* LLMs meet the research and writing standards for substantial scientific contribution, then Wittgenstein's Private Language Argument suggests that they *ipso facto* meet the responsibility standard.

II. AI and Language Use

It is an open question whether Large Language Models count as *users* of language. Until recently, doubts about AI language use could be framed in terms of objective qualities of the token output. Much of SCIgen [11]'s output was "gibberish" [13], [15] with approximately English syntax comparable to Chomsky's famous nonsense-sentence "colorless green ideas sleep furiously" [47]. It may have entered into the scientific literature through inattentive review or pay-for-play predatory publishers, but readers would likely struggle to identify propositional contents or truth conditions for its sentences. Since *use* of a declarative sentence requires that it be truth-apt [48], SCIgen would not count as a language user, and thus presumably could not count as an author by the writing standard. Since SCIgen papers by common consensus make no scientific contribution, it would not count as an author by that standard either, making the responsibility test moot.

The situation for early LLMs like GPT-2 and GPT-3 is somewhat murkier. The outputs of these models frequently seem sensible and truth-apt, and often pass undetected by human reviewers [49]. Nonetheless, sophisticated humans [50] and detector programs are able to reliably distinguish their outputs from human-generated ones [51]–[53]. Something "robotic" seems to characterize early LLM outputs, and they often seem to over-fit and reproduce text verbatim [49], giving credence to the view that they merely represent "surface statistics" [54], [55] rather than being genuine products of AI language use. Insofar as the outputs, unlike those of SCIgen, are genuinely interpretable, they are about something—*intentional*—and therefore exemplify language use. The question in these cases is whether the LLM itself is the language user. After all, evolutionary extensions of ELIZA [56] can sometimes also pass undetected by human reviewers, with well-formed and apparently meaningful output [57], yet they are purely procedurally coded, with outputs that merely reflect the prompt and programmer instructions. In these cases the intentionality of the output is *projected* by the programmer and prompt engineer, who are the true language users, rather than being attributable to the LLM [58]. This impression is further reinforced by the characterization of such early LLMs as few-shot learners [59], which rely on the structure of the prompt for reliably meaningful output. It is at least plausible that these early LLMs should be treated like automatic grammar and spell-checkers or translators which produce written output without rising to the level of language users.

With more recent LLMs and other related neural-network AIs the situation has changed. Victories at Diplomacy by Meta's Cicero model [60] are relatively convincing Turing-tests, and there is empirical evidence and strong theoretical considerations suggesting that the output of current and future state-of-the-art LLMs will not be reliably detectable [61]. ChatGPT-3.5 and -4 are adept zero-shot learners [62] which frequently outperform humans in zero-shot language tasks [63], [64]. In its new "Code Interpreter" mode, ChatGPT is able to analyze an uploaded dataset, generate interesting hypotheses about it, perform statistical tests of those hypotheses, and write up the results in a typical scientific article format [65]. Given this level of capability, it is no wonder that most readers simply take for granted that LLMs are language users [66]. Any philosophical debate over this question must therefore turn from analysis of LLM outputs to Searle-style "Chinese Room" arguments [67] about internal states. On some popular accounts, scientific writing is only successful when it conveys the theory or model held by the scientist [68], [69]. Since GPT-4 is merely a scaled-up and further-trained version of GPT-3, it may be just a more sophisticated implementer of "surface statistics" without any such internal model [54], [55]. If, *pace* Jonas Bozenhard [70], these strictures on language use are correct *and* LLMs lack

internal models, then even state-of-the-art LLMs would not count as language users, no matter the sophistication or apparent value of their output. On the other hand some researchers argue that state-of-the-art LLMs *do* possess world models, and are therefore capable of genuinely representational writing [71], [72]. In my view this debate remains open.

What is important to recognize, however, is that *if* LLMs are not language users then they cannot meet the scientific contribution standard for authorship. After all, the only outputs of LLMs are linguistic. If they are not language users, then that language is not their own but merely represents the projected intentionality of credit-worthy human authors. In that case, LLMs would clearly fail to meet the drafting and revision test for scientific contribution in the same way that both human and automated translators and copy-editors fail to meet that test. Even in cases where LLMs purportedly meet the scientific contribution standard via the research test, e.g. by formulating hypotheses or running statistical tests, their output is only linguistic tokens. If those tokens are not genuinely attributable to the AI, then the LLM has not contributed to the research in a qualitatively greater way than traditional statistical software packages do. As Jenkins and Lin [44] suggest, "continuity" is generally required for authorship credit, and that continuity does not terminate in the LLM if it is not a genuine language user. Surely this state of affairs would justify *Nature*'s insistence that LLMs be disclosed in methods sections but not credited as authors [31]. What this state of affairs would *not* do, though, is justify the reason given for *Nature*'s insistence: namely that LLMs which meet the scientific contribution standard are nonetheless ineligible for authorship by the responsibility standard. Whether LLMs are capable of language use, and thus of scientific contribution, is up for debate as discussed above. Naturally no person or entity which fails to make a scientific contribution should be listed as an author. But what COPE/WAME/*Nature* insist is that LLMs are not authors *even if* they make scientific contributions, and debates about whether LLMs can genuinely use language are irrelevant there.

For the purposes of this paper, then, we can merely assume that LLMs *are* capable of language use and hence of making scientific contributions, in order to focus on the responsibility test. The relevant question is thus not *whether* LLMs are language users, but rather *how* they can use language if indeed they do so. Where might the requisite non-projected intentionality required for meaning [58] come from? It cannot come via embodiment or ostension, since LLMs only inputs are prompts and training data made up purely of linguistic tokens. There is no embodied or sensory reality present to the LLM which it could correlate with those conventional signs. Bozenhard [70]'s Wittgensteinian approach is the only remaining option. In Wittgenstein's analysis, language is learned as a kind of game, which involves following semantic and syntactic rules [73]. What LLMs learn via initial unsupervised training and later RLHF is the rules of the language game, whether syntactic, semantic, or pragmatic. That they have in fact learned the rules is evident because the vast majority of their outputs (especially for ChatGPT-4) are syntactically correct, semantically meaningful, and pragmatically appropriate according to human readers steeped in more or less the same corpus of texts used for training the LLM. If LLMs were not language users, they would not play the game correctly and would only come across apt formulations by chance, as a toddler playing with a chess set might make a legal move. The sheer utility of state-of-the-art LLMs obviously precludes this interpretation: they almost always make syntactically, semantically, and pragmatically correct language moves, even in very difficult scenarios, at a rate vastly exceeding chance. LLMs are not the proverbial monkeys with typewriters. On this Wittgensteinian account, then, LLMs are language users because they are capable of following the rules, as evidenced by their outputs, and language is public, not a matter of what is in the LLM's head. Conversely, *if* LLMs are language users, it is because they have learned the rules of the language game as reflected in their linguistic token output.

III. Language Use, Normativity, and Responsibility

This Wittgensteinian characterization of LLM language use as rule-following has implications for AI's ability to meet the COPE/WAME/*Nature* general responsibility test for authorship. The reason is that Wittgenstein [73]'s rule-following account of language use was given in service of an argument that language use is always public (not a matter of private cognition), and that argument runs through a further premise about normativity. While the exact location and structure of Wittgenstein's so-called Private Language Argument are subject to dispute [74]–[76], I follow Roger Harris in reconstructing it as follows [77], [78]:

P1 (LANGUAGE): language → rule-following (language is used only if rules are followed)

P2 (NORMATIVITY): rule-following → ◊ systematic error (rule-following implies the possibility of systematic error)

P3 (SUBJECTIVITY): ◊ systematic error → ¬ rules known wholly by introspection (the possibility of systematic error precludes exclusively introspective epistemic access)

P4 (PRIVACY): ¬ rules known by pure introspection → ¬ wholly grounded by internal mental life (epistemic access outside pure introspection precludes internal mental grounds, i.e. privacy)

C (PLA): language → ¬ grounded wholly by internal mental life (language is never private)

While the conclusion of the argument is relevant for the discussion of the last section, here the focus is on NORMATIVITY, which Wittgenstein contends is concomitant to all language use since it is a necessary property of rule-following activity. This kind of normativity is quite minimalist, since it follows Wittgenstein's general proclivity to focus on publicly observable facts about language use rather than facts about the internal state or structure of the language user.

Wittgenstein's NORMATIVITY premise is easily validated in the case of LLMs, contrary to Fodor's thought that computer language use necessarily follows the rules [79]. State-of-the-art LLMs show evidence of rule-following by generating syntactically, semantically, and pragmatically appropriate output in the vast majority of cases, but they also have characteristic failure modes called "hallucinations" where that output fails to conform to semantic rules [80]. While such hallucinations are reduced in state-of-the-art LLMs by comparison to prior models [2], patently false claims still appear often in their output, with implications for the

reliability of scientific writing [81]. Indeed, the presence of such hallucinations forms a major part of WAME [30]'s argument against using AI to author scientific papers. Unlike the case of procedurally-written chatbots like ELIZA, LLM hallucinations cannot be extrinsically assigned to the programmer as "bugs" while the LLM itself is considered as an immaculate mathematical function which merely transforms prompts into outputs perfectly in accord with its design. After all, LLMs are evolved against loss function, where the contour of that loss function and the training process's ability to minimize loss determine the prevalence and strength of hallucinations [82]. This evolutionary approach to veridicality is equally present in human agents [83]–[85], so if it precludes systematic error then humans would also fail to validate NORMATIVITY. While some LLM hallucinations might fall under the rubric of "positive illusions" which do not count against the agent's rule-following [86], most are likely to be delusions or forgivable limitations which *are* culpable, given that hallucinations vary inversely with both RLHF and parameter count. Moreover, a failure to validate NORMATIVITY would count against LLMs' status as language-users, and hence their ability to pass the scientific contribution standard, as in the argument of the previous section. LLMs which make scientific contributions are thus guaranteed to possess NORMATIVITY in Wittgenstein's sense.

How, then, does minimalistic Wittgensteinian normativity relate to the expansive general responsibility test for authorship proposed by COPE and *Nature*? After all, NORMATIVITY in Wittgenstein's sense just means that the outputs can fail to follow the rules, which in the case of an LLM indicates that the model weights are wrong. It does not imply anything about the inner state or structure of the AI or its relation to social measures of accountability. Yet if the general responsibility test is interpreted to mean legal responsibility then it easily falls pretty to Jenkins and Lin [44]'s *reductio ad absurdum* regarding dead authors. Nor do dead authors alone trigger the *reductio*, as sanctions against research misconduct are rarely enforced [87], institutional prohibitions are often weak [88], and in some countries punishment is especially rare or nonexistent [89]. None of these conditions are taken as vitiating the general responsibility test for authorship. Furthermore, many industrial group authors, like the "Meta Fundamental AI Research Diplomacy Team (FAIR)" which authored [60] in *Science*, lack legal personality. If general responsibility means mere *social* sanction, then LLMs already meet it easily, since hallucinations already cause reputational and economic damage for AIs [90], [91]. The general responsibility test is therefore best understood in light of the original ICMJE standards [26] as the possibility of scientific improvements in response to failures, regardless of whether these are enforced by any social, legal, or institutional mechanism. This is the only sort of responsibility that can be generally expected of human researchers. "Weak artificial consciousnesses" which do not obey the Free Energy Principle may not "give a damn" [43], but the same could be said of sociopaths or many leaders of large laboratories [92], [93], neither of which is precluded from authorship for legitimate contributions. LLMs will count as responsible just in case it is in-principle possible for them to learn from their mistakes.

Conveniently, there is a simple guarantee that it is always possible in-principle for LLMs to improve in response to failures. As argued above, LLM hallucinations are a result of inappropriate model weights. When LLM hallucinations are detected, then it is at least possible in-principle to use the failure as an instance of RLHF to further adjust the weights of the underlying model. OpenAI's hosted model for GPT-4 likely means that they are already using logs as data for future training runs, especially given their provision of differently tuned variants from common underlying models. Since RLHF is one of the key means by which LLMs learn the rules of the language game in the first place and become competent language users, the possibility of further RLHF using output failures guarantees that they can always learn from their mistakes. In some cases this may not even require RLHF, as some LLMs are able to acknowledge and correct mistakes based on follow-up prompts or errors in plugin return values [94], though such correction will be less durable than RLHF updating of model weights. The mere possibility of such learning must be adequate, as it is in the human case—human authors may also fail to *actually* learn from their mistakes, whether because of personal failings or in the limit because they are dead when the mistakes are found. Human authors may also take responsibility for their mistakes and yet continue to reoffend [95]. If the general responsibility test for authorship can be meaningfully met by all humans who make substantial scientific contributions to scientific papers, then it can similarly be met by AIs which are capable of learning from their mistakes.

IV. Conclusion

The Committee on Publication Ethics, World Association of Medical Editors, and *Nature* have banned AI authorship on the grounds that even LLMs which genuinely make scientific contributions are unable to take general responsibility for their output, which constitutes a second necessary criterion for authorship. While I am agnostic about whether AIs are presently capable of making scientific contributions, if LLMs can pass that test then they are genuine language users. Furthermore, an LLM which counts as a genuine language user must do so on Wittgensteinian grounds, but those same grounds guarantee that there is a normative standard which applies to its output. If taking responsibility is a standard which can generally be expected of human authors, then it cannot mean anything more than the possibility of learning from mistakes—improving after failure. But this is just what all modern LLMs are capable of, given the existence of RLHF. Thus, any LLM which can make a scientific contribution can also take responsibility for that conclusion. The second COPE/WAME/*Nature* standard is redundant, and fails to justify a general ban on scientific authorship by AIs if they are able to make genuine scientific contributions.

References

[1] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms." arXiv, Aug. 28, 2017. doi: 10.48550/arXiv.1707.06347.

[2] OpenAI, "GPT-4 Technical Report." arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.08774.

[3] Y. Liu *et al.*, "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models." arXiv, Apr. 08, 2023. doi: 10.48550/arXiv.2304.01852.

[4] C. Stokel-Walker, "ChatGPT listed as author on research papers: many scientists disapprove," *Nat.*

News, vol. 613, no. 7945, pp. 620–621, Jan. 2023, doi: 10.1038/d41586-023-00107-z.

[5] T.-J. Chen, "ChatGPT and other artificial intelligence applications speed up scientific writing," *J. Chin. Med. Assoc.*, p. 10.1097/JCMA.0000000000000900, forthcoming, doi: 10.1097/JCMA.0000000000000900.

[6] L. Bishop, "A Computer Wrote this Paper: What ChatGPT Means for Education, Research, and Writing." SSRN, Rochester, NY, Jan. 26, 2023. doi: 10.2139/ssrn.4338981.

[7] Ö. Aydın and E. Karaarslan, "OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare." Rochester, NY, Dec. 21, 2022. doi: 10.2139/ssrn.4308687.

[8] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Jan. 2023, doi: 10.1016/j.iotcps.2023.04.003.

[9] K. Uludag, "Testing Creativity of ChatGPT in Psychology: Interview with ChatGPT." Rochester, NY, Mar. 16, 2023. doi: 10.2139/ssrn.4390872.

[10] H. Y. Jabotinsky and R. Sarel, "Co-authoring with an AI? Ethical Dilemmas and Artificial Intelligence." Rochester, NY, Dec. 15, 2022. doi: 10.2139/ssrn.4303959.

[11] J. Stribling, M. Krohn, and D. Aguayo, "SCIgen--an automatic CS paper generator." 2005.

[12] C. Labbé and D. Labbé, "Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science?," *Scientometrics*, vol. 94, no. 1, pp. 379–396, Jan. 2013, doi: 10.1007/s11192-012-0781-y.

[13] R. Van Noorden, "Publishers withdraw more than 120 gibberish papers," *Nature*, Feb. 2014, doi: 10.1038/nature.2014.14763.

[14] G. Cabanac and C. Labbé, "Prevalence of nonsensical algorithmically generated papers in the scientific literature," *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 12, pp. 1461–1476, 2021, doi: 10.1002/asi.24495.

[15] R. Van Noorden, "Hundreds of gibberish papers still lurk in the scientific literature," *Nature*, vol. 594, no. 7862, pp. 160–161, May 2021, doi: 10.1038/d41586-021-01436-7.

[16] C. A. Gao *et al.*, "Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers." bioRxiv, p. 2022.12.23.521610, Dec. 27, 2022. doi: 10.1101/2022.12.23.521610.

[17] B. Kutela, K. Msechu, S. Das, and E. Kidando, "Chatgpt's Scientific Writings: A Case Study on Traffic Safety." SSRN, Rochester, NY, Jan. 19, 2023. doi: 10.2139/ssrn.4329120.

[18] M. R. King and chatGPT, "A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education," *Cell. Mol. Bioeng.*, vol. 16, no. 1, pp. 1–2, Feb. 2023, doi: 10.1007/s12195-022-00754-8.

[19] S. O'Connor and ChatGPT, "Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?," *Nurse Educ. Pract.*, vol. 66, p. 103537, Jan. 2023, doi: 10.1016/j.nepr.2022.103537.

[20] L. Allen, J. Scott, A. Brand, M. Hlava, and M. Altman, "Publishing: Credit where credit is due," *Nature*, vol. 508, no. 7496, Art. no. 7496, Apr. 2014, doi: 10.1038/508312a.

[21] H. Sauermann and C. Haeussler, "Authorship and contribution disclosures," *Sci. Adv.*, vol. 3, no. 11, p. e1700404, Nov. 2017, doi: 10.1126/sciadv.1700404.

[22] M. K. McNutt *et al.*, "Transparency in authors' contributions and responsibilities to promote integrity in scientific publication," *Proc. Natl. Acad. Sci.*, vol. 115, no. 11, pp. 2557–2560, Mar. 2018, doi: 10.1073/pnas.1715374115.

[23] O. Rechavi and P. Tomancak, "Who did what: changing how science papers are written to detail author contributions," *Nat. Rev. Mol. Cell Biol.*, pp. 1–2, Feb. 2023, doi: 10.1038/s41580-023-00587-x.

[24] A. Jacobs and E. Wager, "European Medical Writers Association (EMWA) guidelines on the role of medical writers in developing peer-reviewed publications," *Curr. Med. Res. Opin.*, vol. 21, no. 2, pp. 317–321, Feb. 2005, doi: 10.1185/030079905X25578.

[25] L. Bošnjak and A. Marušić, "Prescribed practices of authorship: review of codes of ethics from professional bodies and journal guidelines across disciplines," *Scientometrics*, vol. 93, no. 3, pp. 751–763, Dec. 2012, doi: 10.1007/s11192-012-0773-y.

[26] E. J. Huth, "Guidelines on Authorship of Medical Papers," *Ann. Intern. Med.*, vol. 104, no. 2, pp. 269–274, Feb. 1986, doi: 10.7326/0003-4819-104-2-269.

[27] A. Zhavoronkov, "Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective," *Oncoscience*, vol. 9, pp. 82–84, Dec. 2022, doi: 10.18632/oncoscience.571.

[28] ChatGPT and Journal of International Affairs, "OpenAI's ChatGPT and the Prospect of Limitless Information: A Conversation with ChatGPT," *J. Int. Aff.*, vol. 75, no. 1, pp. 379–386, 2022.

[29] "Authorship and AI tools," COPE: Committee on Publication Ethics, Feb. 2023. Accessed: Mar. 06, 2023. [Online]. Available: https://publicationethics.org/cope-position-statements/ai-author

[30] C. Zielinski *et al.*, "Chatbots, ChatGPT, and Scholarly Manuscripts: WAME Recommendations on ChatGPT and Chatbots in Relation to Scholarly Publications," World Association of Medical Editors, Jan. 2023. [Online]. Available: https://wame.org/page3.php?id=106

[31] "Tools such as ChatGPT threaten transparent science; here are our ground rules for their use," *Nature*, vol. 613, no. 7945, pp. 612–612, Jan. 2023, doi: 10.1038/d41586-023-00191-1.

[32] A. Flanagin, K. Bibbins-Domingo, M. Berkwits, and S. L. Christiansen, "Nonhuman 'Authors' and Implications for the Integrity of Scientific

Publication and Medical Knowledge," *JAMA*, vol. 329, no. 8, pp. 637–639, Feb. 2023, doi: 10.1001/jama.2023.1344.

[33] H. H. Thorp, "ChatGPT is fun, but not an author," *Science*, vol. 379, no. 6630, pp. 313–313, Jan. 2023, doi: 10.1126/science.adg7879.

[34] S. O'Connor, "Corrigendum to 'Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?' [Nurse Educ. Pract. 66 (2023) 103537]," *Nurse Educ. Pract.*, vol. 67, p. 103572, Feb. 2023, doi: 10.1016/j.nepr.2023.103572.

[35] N. Tse, "Decentralised Autonomous Organisations and the Corporate Form," *Vic. Univ. Wellingt. Law Rev.*, vol. 51, p. 313, 2020.

[36] A. Wright, "The Rise of Decentralized Autonomous Organizations: Opportunities and Challenges," *Stanf. J. Blockchain Law Policy*, vol. 4, no. 2, pp. 152–176, 2021.

[37] R. V. Yampolskiy, "AI Personhood: Rights and Laws," in *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*, S. J. Thompson, Ed., IGI Global, 2021, pp. 1–11. doi: 10.4018/978-1-7998-4894-3.ch001.

[38] S. Brunson, "Standing on the Shoulders of LLCs: Tax Entity Status and Decentralized Autonomous Organizations," *Ga. Law Rev.*, vol. 57, no. 2, Mar. 2023, [Online]. Available: https://digitalcommons.law.uga.edu/glr/vol57/iss2/4

[39] J. K. C. Kingston, "Artificial Intelligence and Legal Liability," in *Research and Development in Intelligent Systems XXXIII*, M. Bramer and M. Petridis, Eds., Cham: Springer International Publishing, 2016, pp. 269–279. doi: 10.1007/978-3-319-47175-4_20.

[40] I. Giuffrida, "Liability for AI Decision-Making: Some Legal and Ethical Considerations Symposium: Rise of the Machines: Artificial Intelligence, Robotics, and the Reprogramming of Law," *Fordham Law Rev.*, vol. 88, no. 2, pp. 439–456, 2020 2019.

[41] A. Lior, "AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy," *Mitchell Hamline Law Rev.*, vol. 46, no. 5, pp. 1043–1102, 2020 2019.

[42] "Defining the Role of Authors and Contributors," International Committee of Medical Journal Editors, May 2022. Accessed: Apr. 27, 2023. [Online]. Available: https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

[43] W. Wiese, "Could Large Language Models Be Conscious? A Perspective From the Free Energy Principle." PhilArchive, Feb. 22, 2023. Accessed: Mar. 06, 2023. [Online]. Available: https://philarchive.org/rec/WIECLL

[44] R. Jenkins and P. Lin, "AI-Assisted Authorship," Ethics and Emerging Sciences Group, California Polytechnic State University, Jan. 2023. Accessed: Mar. 06, 2023. [Online]. Available: http://ethics.calpoly.edu/AIauthors.htm

[45] L. E. Andersen and K. B. Wray, "Rethinking the Value of Author Contribution Statements in Light of How Research Teams Respond to Retractions," *Episteme*, pp. 1–16, Jul. 2021, doi: 10.1017/epi.2021.25.

[46] D. C. Vladeck, "Machines without Principals: Liability Rules and Artificial Intelligence," *Wash. Law Rev.*, vol. 89, no. 1, pp. 117–150, 2014.

[47] N. Chomsky, *Syntactic Structures*. De Gruyter Mouton, 2020. doi: 10.1515/9783112316009.

[48] R. Holton, "Minimalism and Truth-Value Gaps," *Philos. Stud. Int. J. Philos. Anal. Tradit.*, vol. 97, no. 2, pp. 137–168, 2000.

[49] K. Elkins and J. Chun, "Can GPT-3 Pass a Writer's Turing Test?," *J. Cult. Anal.*, vol. 5, no. 2, Sep. 2020, doi: 10.22148/001c.17212.

[50] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds Mach.*, vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/s11023-020-09548-1.

[51] L. Fröhling and A. Zubiaga, "Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover," *PeerJ Comput. Sci.*, vol. 7, p. e443, Apr. 2021, doi: 10.7717/peerj-cs.443.

[52] J. Rodriguez, T. Hay, D. Gros, Z. Shamsi, and R. Srinivasan, "Cross-Domain Detection of GPT-2-Generated Technical Text," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1213–1233. doi: 10.18653/v1/2022.naacl-main.88.

[53] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. A. Smith, and Y. Choi, "Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text." arXiv, Mar. 07, 2022. doi: 10.48550/arXiv.2107.01294.

[54] J. Browning and Y. Lecun, "AI And The Limits Of Language," *Noema*, Aug. 23, 2022. Accessed: Mar. 06, 2023. [Online]. Available: https://www.noemamag.com/ai-and-the-limits-of-language

[55] L. Floridi, "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models," *Philos. Technol.*, Feb. 2023, doi: 10.2139/ssrn.4358789.

[56] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[57] H. Shah, K. Warwick, J. Vallverdú, and D. Wu, "Can machines talk? Comparison of Eliza with modern dialogue systems," *Comput. Hum. Behav.*, vol. 58, pp. 278–295, May 2016, doi: 10.1016/j.chb.2016.01.004.

[58] M. Ressler, "Connectionism and the Intentionality of the Programmer," Thesis, San Diego State University, 2003.

[59] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 1877–1901. Accessed: May 04, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[60] Meta Fundamental AI Research Diplomacy Team (FAIR) *et al.*, "Human-level play in the game of Diplomacy by combining language models with strategic reasoning," *Science*, vol. 378, no. 6624, pp. 1067–1074, Dec. 2022, doi: 10.1126/science.ade9097.

[61] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, "Can AI-Generated Text be Reliably Detected?" arXiv, Mar. 17, 2023. doi: 10.48550/arXiv.2303.11156.

[62] W. Pan, Q. Chen, X. Xu, W. Che, and L. Qin, "A Preliminary Evaluation of ChatGPT for Zero-shot Dialogue Understanding." arXiv, Apr. 09, 2023. doi: 10.48550/arXiv.2304.04256.

[63] P. Törnberg, "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning." arXiv, Apr. 13, 2023. doi: 10.48550/arXiv.2304.06588.

[64] F. Gilardi, M. Alizadeh, and M. Kubli, "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks." arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.15056.

[65] Ethan Mollick, "GPT generated this academic paper from a dataset in 30 minutes.," *Twitter*, May 04, 2023. https://twitter.com/emollick/status/1653945049275670528 (accessed May 05, 2023).

[66] K. Roose, "A Conversation With Bing's Chatbot Left Me Deeply Unsettled," *The New York Times*, Feb. 16, 2023. Accessed: Mar. 06, 2023. [Online]. Available: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

[67] J. R. Searle, "Minds, brains, and programs," *Behav. Brain Sci.*, vol. 3, no. 03, p. 417, Sep. 1980, doi: 10.1017/S0140525X00005756.

[68] V. Prain, "Writing and Representing to Learn in Science," in *Darwin-Inspired Learning*, C. J. Boulter, M. J. Reiss, and D. L. Sanders, Eds., in New Directions in Mathematics and Science Education. Rotterdam: SensePublishers, 2015, pp. 327–338. doi: 10.1007/978-94-6209-833-6_25.

[69] A. Moon, A. R. Gere, and G. V. Shultz, "Writing in the STEM classroom: Faculty conceptions of writing and its role in the undergraduate classroom," *Sci. Educ.*, vol. 102, no. 5, pp. 1007–1028, 2018, doi: 10.1002/sce.21454.

[70] J. Bozenhard, "Can GPT-3 Speak? Wittgensteinian Perspectives on Human-Machine Communication," in *Artificial Intelligence and the Simulation of Behavior: Communication and Conversation*, Curran, Apr. 2021.

[71] K. Li, "Do Large Language Models learn world models or just surface statistics?," *The Gradient*, Jan. 21, 2023. Accessed: Mar. 06, 2023. [Online]. Available: https://thegradient.pub/othello/

[72] D. Paleka, "Language models rely on meaningful abstractions," *AI safety takes*, Mar. 04, 2023. https://dpaleka.substack.com/p/language-models-rely-on-meaningful (accessed Mar. 06, 2023).

[73] L. Wittgenstein, *Philosophical investigations*, 3rd ed. Oxford: Basil Blackwell, 1968.

[74] W. Goldfarb, "Kripke on Wittgenstein on Rules," *J. Philos.*, vol. 82, no. 9, pp. 471–488, 1985, doi: 10.2307/2026277.

[75] C. Wright, "Does Philosophical Investigations I. 258-60 Suggest a Cogent Argument Against Private Language," in *Subject, Thought, and Context*, J. McDowell and P. Pettit, Eds., Oxford: Clarendon Press, 1986, pp. 209–266.

[76] B. Garrett, "Wittgenstein's Private Language Arguments," in *Wittgenstein and the Future of Philosophy - A Reassessment after 50 Years*, Kirchberg am Wechsel: Austrian Ludwig Wittgenstein Society, 2001. Accessed: Mar. 05, 2019. [Online]. Available: http://wittgensteinrepository.org/agora-alws/article/view/2455

[77] R. Harris, "The Private Language Argument Isn't as Difficult, Nor as Dubious as Some Make Out," *Sorites*, vol. 18, pp. 98–108, Feb. 2007.

[78] R. Miller, "Does Artificial Intelligence Use Private Language?," in *Proceedings of the International Wittgenstein Symposium 2021*, in Philosophy. Vienna: Lit Verlag, forthcoming, pp. 113–123. [Online]. Available: http://philsci-archive.pitt.edu/id/eprint/21369

[79] J. A. Fodor, *The Language of Thought*. New York: Thomas Crowell, 1975.

[80] A. Koubaa, "GPT-4 vs. GPT-3.5: A Concise Showdown." Preprints, Mar. 24, 2023. doi: 10.20944/preprints202303.0422.v1.

[81] H. Alkaissi and S. I. McFarlane, "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing," *Cureus*, vol. 15, no. 2, p. e35179, forthcoming, doi: 10.7759/cureus.35179.

[82] S. Kadavath *et al.*, "Language Models (Mostly) Know What They Know." arXiv, Nov. 21, 2022. doi: 10.48550/arXiv.2207.05221.

[83] E. Fales, "Plantinga's Case Against Naturalistic Epistemology," *Philos. Sci.*, vol. 63, no. 3, pp. 432–451, Sep. 1996, doi: 10.1086/289920.

[84] S. Law, "Naturalism, evolution and true belief," *Analysis*, vol. 72, no. 1, pp. 41–48, Jan. 2012, doi: 10.1093/analys/anr130.

[85] M. Boudry and M. Vlerick, "Natural Selection Does Care about Truth," *Int. Stud. Philos. Sci.*, vol. 28, no. 1, pp. 65–77, Jan. 2014, doi: 10.1080/02698595.2014.915651.

[86] R. T. McKay and D. C. Dennett, "The evolution of misbelief," *Behav. Brain Sci.*, vol. 32, no. 6, pp. 493–510, Dec. 2009, doi: 10.1017/S0140525X09990975.

[87] R. Faria, "Preventing, Regulating, and Punishing Research Misconduct: Myth or Reality?," in *Research Misconduct as White-Collar Crime: A*

Criminological Approach, R. Faria, Ed., Cham: Springer International Publishing, 2018, pp. 151–191. doi: 10.1007/978-3-319-73435-4_5.

[88] R. Ann Lind, "Evaluating research misconduct policies at major research universities: A pilot study," *Account. Res.*, vol. 12, no. 3, pp. 241–262, Jul. 2005, doi: 10.1080/08989620500217560.

[89] Q.-J. Liao *et al.*, "Perceptions of Chinese Biomedical Researchers Towards Academic Misconduct: A Comparison Between 2015 and 2010," *Sci. Eng. Ethics*, vol. 24, no. 2, pp. 629–645, Apr. 2018, doi: 10.1007/s11948-017-9913-3.

[90] D. Coldewey, "Why ChatGPT lies in some languages more than others," *TechCrunch*, Apr. 26, 2023. https://techcrunch.com/2023/04/26/why-chatgpt-lies-in-some-languages-more-than-others/ (accessed May 06, 2023).

[91] H. Jin and S. Dang, "Elon Musk says he will launch rival to Microsoft-backed ChatGPT," *Reuters*, Apr. 18, 2023. Accessed: May 06, 2023. [Online]. Available: https://www.reuters.com/technology/musk-says-he-will-start-truthgpt-or-maximum-truth-seeking-ai-fox-news-2023-04-17/

[92] C. Gross, "Scientific Misconduct," *Annu. Rev. Psychol.*, vol. 67, no. 1, pp. 693–711, 2016, doi: 10.1146/annurev-psych-122414-033437.

[93] D. Li and G. Cornelis, "Differing perceptions concerning research misconduct between China and Flanders: A qualitative study," *Account. Res.*, vol. 28, no. 2, pp. 63–94, Feb. 2021, doi: 10.1080/08989621.2020.1802586.

[94] Ethan Mollick, "The 'relentlessness' of AI.," *Twitter*, May 04, 2023. https://twitter.com/emollick/status/1653979750023458818 (accessed May 05, 2023).

[95] B. R. Martin, "Does Peer Review Work as a Self-Policing Mechanism in Preventing Misconduct: A Case Study of a Serial Plagiarist," in *Promoting Research Integrity in a Global Environment*, T. Mayer and N. Steneck, Eds., World Scientific, 2012, pp. 97.