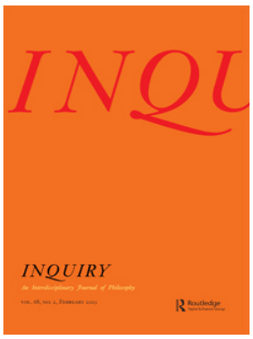

Routledge
Taylor & Francis Group

Inquiry
An Interdisciplinary Journal of Philosophy

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/sinq20

Publishing robots

Nicholas Hadsell, Rich Eva & Kyle Huitt

To cite this article: Nicholas Hadsell, Rich Eva & Kyle Huitt (11 Feb 2025): Publishing robots, Inquiry, DOI: 10.1080/0020174X.2025.2460764

To link to this article: https://doi.org/10.1080/0020174X.2025.2460764

INQUIRY
https://doi.org/10.1080/0020174X.2025.2460764

Routledge
Taylor & Francis Group

Publishing robots

Nicholas Hadsell[a], Rich Eva[b] and Kyle Huitt[a]

[a]Philosophy Department, Baylor University, Waco, TX, United States; [b]Pratt School of Engineering, Duke University, Durham, NC, United States

ABSTRACT
If AI can write an excellent philosophy paper, we argue that philosophy journals should strongly consider publishing that paper. After all, AI stands to make significant contributions to ongoing projects in some subfields, and it benefits the world of philosophy for those contributions to be published in journals, the primary purpose of which is to disseminate significant contributions to philosophy. We also propose the *Sponsorship Model* of AI journal to mitigate any costs associated with our view. This model requires at least one human to sponsor any AI manuscript submitted to a journal. After making this case, we evaluate a variety of objections – that our proposal will crash the academic job market, disincentivize humans from doing philosophy, and violate a variety of justified publishing norms – and conclude that they are all wanting.

ARTICLE HISTORY Received 15 May 2024; Accepted 26 January 2025

KEYWORDS Ethics of AI; research ethics; technology ethics

Imagine you are a journal referee who just read an excellent article. The submission accurately summarizes the existing literature, identifies a problem, and resolves that problem with a unique and compelling solution. Moreover, it is superbly written; the prose is comparable to that of David Lewis, Bertrand Russell, and Judith Jarvis Thomson. You feel both honored and underqualified to referee the article. In normal circumstances, you would unconditionally recommend this article for acceptance. But this is no normal circumstance: You have just found out the submission was entirely generated by artificial intelligence. Should you change your recommendation? We say you should *not* change your recommendation.

This scenario may strike some readers as a far-off fiction, but it is closer than you think. AI has already made its way into all kinds of industries.[1]

CONTACT Nicholas Hadsell Nicholas_Hadsell1@baylor.edu 1311 S 5th St, Waco, TX 76706, United States

[1]All pre-2019 citations in this paragraph come from Danaher 2019.

In finance, AI conducts at least half of all trades in the United States (Wellman and Rajan 2017). AI will probably do all the medical diagnostic work very soon (Lysandrou et al. 2023) and is already starting to do some of the caretaking work (Ayers et al. 2023). AI lawyers research and assist human lawyers in their cases (Koebler 2017; Martin et al. 2024). The academic world is not immune to AI intervention either. Logicians are developing AI that can generate and test their proofs (Harris 2015), and some AI can conduct scientific experiments to test hypotheses (Sparkes et al. 2010; Williams et al. 2015).

We might think the art of philosophical writing is beyond the reach of AI, but we would be wrong. Schwitzgebel et al. trained an AI on the work of Daniel Dennett, and experts on Dennett's work had difficulty distinguishing between Dennett's actual words and the AI-generated ones (Schwitzgebel, Schwitzgebel, and Strasser 2023). AI can already emulate analytic philosophers in a passable and compelling way. It is only a matter of time before large language models develop their unique voices and contributions to academic philosophy. What should we do once this happens?

In our view, if AI can write an excellent philosophy paper, philosophy journals should strongly consider publishing that paper. After all, the main purpose of philosophy journals is the dissemination of excellent philosophy papers. If AI can meet this purpose, we see little reason to exclude them from journals.

We should define two terms at the outset. First, by an 'excellent paper', we mean a paper that is original (not plagiarized), well-written, timely, compelling, and insightful. It is debated whether AI can produce such work, especially whether it can produce anything original. We think that this is possible. Under the right prompting, for example, it seems that AI could write an appropriately cited, original meta-analysis of a previously unanalyzed area of philosophy. We could be mistaken, but this is at least an open question. So, our aim in this paper is to move beyond questions about AI's theoretical capabilities to ask 'What should we do once AI has the capability to produce original work?' In other words, we are tabling the plagiarism question. Let us assume, for the sake of argument, that AI can write an 'excellent paper'.[2] Second, by 'strongly

[2]We acknowledge many large language models plagiarize (Lee et al. 2023). Our argument is conditional on the assumption that AI could produce non-plagiarized papers. If AI can do this now, our argument is relevant. If AI is only currently incapable of producing non-plagiarized work but could in the future, then our argument becomes relevant in the future. Of course, if AI is, in principle, incapable of producing non-plagiarized work, then this is a trivial thought experiment. For our purposes here, we bracket this last possibility.

consider', we mean that journals should primarily assess work based on its excellence, not on its authorship. This would mean that AI articles would be openly considered and likely published by many journals. We say 'strongly consider' because we want to leave open the option for journals to publish human-only volumes, and even for some journals to remain exclusive to human authors, much like some journals are exclusive to undergraduates.

In this paper, we will also propose a practical way journals can begin accepting AI work that minimizes some of the potential costs of our proposal. Our preferred method is the *Sponsorship Model*. In this model, journal editors require AI work to have a human author who sponsors the article. The Sponsorship Model will help editors avoid infinitely large submission queues, evaluate articles more efficiently, and add another layer of accountability to the publication process.

In part one, we consider the purpose of academic philosophy journals. In part two, we survey how AI-generated journal articles could substantially benefit academic philosophers. In part three, we introduce the Sponsorship Model. In part four, we evaluate objections to our proposal and find them wanting. We conclude that once AI can produce excellent work, philosophy journals should strongly consider that work for publication.

1. What are journals for?

Before we get to our proposal, we should ask: What is the purpose of academic publishing? We want to suggest a *main* purpose that most people would agree with. Put simply, it is to disseminate excellently written philosophy. Publishing may have ancillary purposes, and this definition could be further analyzed and expanded, but something like this purpose seems essential to publishing. When graduate students write an excellent paper, they are advised to send it to a journal. When you write an excellent article, you feel like you ought to send it to a journal. Your article and the graduate student's article could be published on a personal blog or some other venue, but we tend to think that journals are *for* excellent articles. When reviewers and editors consider a submission, they ask themselves something like, 'Is this an excellent work of philosophy?' If not, the journal will, we hope, reject the article. After all, top generalist philosophy journals typically endorse this view in their mission statements. *Noûs* and *Philosophy and Phenomenological Research* both claim their

mission is to '[publish] articles that make standalone, substantial contributions'.[3] *Mind* aspires to publish 'cutting edge philosophical papers' and 'the best philosophy in a variety of fields'.[4] *Philosophical Studies* states that their 'principal aim is to publish articles that are models of clarity and precision in dealing with significant philosophical issues'.[5]

To demonstrate that this purpose is widely acknowledged, consider this case:

Sleepwriting: Bob wakes up in the middle of the night to find that he wrote an excellent philosophical paper in his sleep. He sends the paper to a journal with a cover letter explaining his sleepwriting miracle. The reviewers and editors unanimously accept the paper before reading his cover letter.

Should the editorial team publish Bob's paper? We suspect most readers will intuitively answer *yes*. After all, why should it matter if Bob wrote the paper consciously or unconsciously? He was honest about how the paper was written. The editorial team should publish the paper because it is excellent. Bob's writing process could have been very different: Maybe he wrote the paper by pressing his keyboard randomly for thirty minutes straight. Perhaps he tossed his laptop down a flight of stairs and, voilà, the laptop generates the same excellent paper. In virtually all cases, we think the editorial team should publish the paper regardless of how it came about.

Of course, one might think publishing has another main purpose: to be a heuristic for effective job candidates. This is a reasonable idea; hiring committees use research records to evaluate candidates. But there may be cases where this purpose clashes with what we think the *main* purpose of publishing is, and it is clear that the main purpose wins out. For example, what if Bob makes a career out of sleepwriting papers in top journals despite being a terrible philosopher while awake? Consider this case:

Sleepwriter on the Market: Bob and Maria are the two finalists for a position. Their CVs are equally excellent, except Bob has a slightly better research record (that he produced in his sleep) than Maria.

We might worry that *Sleepwriter on the Market* shows how Bob could undermine hiring committees' abilities to use research records as a hiring heuristic. If publishing is all about the dissemination of excellent

[3]https://onlinelibrary.wiley.com/journal/14680068; https://onlinelibrary.wiley.com/journal/19331592.
[4]https://academic.oup.com/mind/pages/About.
[5]Aims and scope | Philosophical Studies (springer.com).

philosophy, Bob might be able to trick his peers into thinking he is a better researcher than he is.[6]

However, this is a worry *not* about the primary purpose of academic journals but the effectiveness of job searches. Surely, in hiring contexts, committees will use academic publishing as a heuristic for determining whether a candidate is a good researcher. However, this case does not undermine our claim that the primary purpose of journals is to disseminate excellent philosophy; it just shows that sleepwriting Bob might have an advantage over his competitors. Even though committees will use academic publishing as a heuristic for finding suitable candidates, this does not say anything about the purpose of journals *per se*. After all, editorial evaluations of journal submissions are not sensitive in the slightest toward how the submission they are evaluating would mislead hiring committees or advance an author's career. Consider this case:

> *The Undergraduate Prodigy*: Nikita is an intelligent undergraduate who submitted her Intro to Philosophy term paper to a top journal. All reviewers and editors unanimously accepted the paper. The editorial team then discovered that Nikita's career prospects in academic philosophy will skyrocket if this paper is published in their journal.

Should the editorial team rescind their decision after learning that Nikita is an undergraduate who may or may not go into academia? Or, in the other direction, should they count Nikita's status as even more reason to accept? Of course not. The editorial team should consider the submission only on its merits. Factors like career prospects are not germane to peer review, and that is partially why submissions are anonymous. Maybe a publication will help one hiring committee find a successful candidate, or maybe it will mislead another committee into thinking someone is a good researcher when they are just lucky. But considerations like these are beyond the purview of journal editorial teams; they just need to know if the paper is excellent.[7] Hiring committees are responsible for figuring out the rest.[8]

[6]However, *Sleepwalker on the Market* does not obviously indicate Bob's research output is unfit for the job. As long as Bob's papers meet editorial standards, he generates research. Even though his writing process is a bit unconventional – i.e. he tries to go to sleep with the hope that he wakes up in front of an excellent paper – it churns out research nonetheless.

[7]Here is another reason to reject the idea that the purpose of academic publishing is for career evaluation: Imagine a utopian world where no one needs to work and everyone's needs are met. Would we still have something akin to academic publishing? We think so. But if there are no academic institutions – no hiring committees using journals as a hiring heuristic – then that is good evidence the purpose of journals goes beyond such institutions.

[8]While the main purpose of publishing is to disseminate excellent work, there are exceptions. For example, there are possible cases when an editorial team should not publish an otherwise excellent

2. Benefits

Now that we have an idea of what the main purpose of academic publishing is, we can ask how AI could help us meet that purpose. Consider how AI has revolutionized another sort of human activity: chess. About 26 years after Garry Kasparov first lost to a programd chess computer, AI has learned to play chess far better than any human can, and this has led to massive revolutions and advancements in chess theory, enabling the top players to play more precisely than ever before. Chess has had a resurgence in popularity. AI chess engines did not ruin chess for humans; they refined and improved it.[9] We think something similar will happen for philosophy.

Imagine AI generates an excellent paper about Aristotelian metaphysics. If read by those who work in that area, it is the kind of paper that would advance discussion on the key problems. A philosopher who had just been messing around with the AI when she asked it for a paper on Aristotelian metaphysics reads the paper, realizes how philosophically insightful it is, and sends it off to a journal (while letting the editorial team know the paper is AI-generated, of course). If the journal accepts the paper, the expected benefits are straightforward enough: Progress will have been made on an important philosophical issue; insights will have been shared and gained; we will better understand an important, deep issue; those who work on that issue will be able to make progress they otherwise could not have achieved on their own.[10] And so on.

The benefits above would result from any excellent philosophy article. But we expect AI to be especially good at providing certain kinds of insights. We think, for example, that AI will become very good at identifying when authors are talking past one another, ignoring important theoretical possibilities, or are unaware of important work that has consequences for their own published ideas (including work from other fields). If AI can identify and articulate these problems in the literature, it would make welcome contributions to journals that will help

journal article because the consequences of doing so are too severe: if the author turns out to be a famous serial killer, if the findings in the paper were acquired in nefarious ways, if the paper will incite violence against someone, etc. Exceptions will exist for any non-absolutist principle; there will always be all-things-considered reasons to go against the principle that otherwise should guide our conduct in most cases. Whatever all-things-considered threshold there is for our claim, we doubt most plausible scenarios would cross it.

[9]See Webb 2023 for an excellent discussion about how chess players have used AI to improve their game. AlphaGo is another AI that has helped players train in the game of Go. See Silver et al. 2016.

[10]GPT-4 is already able to produce unique responses to prompts that are hard to distinguish from those of humans (Guo et al. 2023; found in Clay and Ontiveros 2023: 466).

philosophers move forward.[11] We also look forward to AI being able to do comprehensive meta-analyses of literature that would be far too time-consuming left to humans. Many disciplines use meta-analyses to help researchers grasp work already done in some area. Once AI is sufficiently developed, journals could leverage it to publish meta-analyses that can help academics get on the same page and move conversations forward more efficiently. Moreover, we think it is probable that developers will be able to train AI to identify how and when philosophers introduce new concepts or identify new problems, allowing AI to make similar contributions. An AI that can do this would be a powerful asset for helping philosophers avoid groupthink and approach things in new ways they might have yet to develop themselves. Finally, AI has the potential to be very useful in more formal areas of philosophy where there are well-defined rules and goals. AI can tell a chess player when there is a position that can force a checkmate, and AI stands to be able to tell philosophers when it has achieved a proof relevant to some issue that philosophers are working on. Researchers have already leveraged AI to generate novel proofs in mathematics that humans could not have developed without computer assistance,[12] And we would welcome advancements from AI in confirmation theory, decision theory, and other formal areas of philosophy.[13]

To summarize, the primary benefit of accepting excellent work done by AI is the same as accepting any excellent work: we can move conversations on philosophical problems forward. More specifically, we think AI will be able to detect and point out ongoing problems in existing literature, take far more data and research into account when analyzing and responding to work that has already been done, help humans approach and introduce concepts in novel and illuminating ways, and help humans get a grasp of vast literature that they could not have achieved on their own. All of this frees up time and energy for philosophers to pursue other important philosophical work like expanding their own expertise, teaching, and bringing philosophy to the general public. A future where AI publishes excellent journal articles is bright.

[11]See Ontiveros and Clay 2023: 466–467 for a good discussion about how AI can plausibly synthesize large literature and offer philosophers recommendations on navigating them.

[12]See Harris 2015 for a discussion about how Timothy Gowers teamed up with computer scientists to use AI to solve various mathematical problems.

[13]See, for example, 'In New Math Proofs, Artificial Intelligence Plays to Win'. https://www.quantamagazine.org/in-new-math-proofs-artificial-intelligence-plays-to-win-20220307/.

3. Our proposal: the sponsorship model

Before we consider objections to publishing AI articles, we want to suggest one method journals should use to if they accept our proposal: the Sponsorship Model. This model curtails some of the costs that may come with AI articles. The motivating principle behind this model is that excellent philosophy articles are good for humans, so humans should remain in the publication process.

The Sponsorship Model requires that all AI-generated work be sponsored by a human or a committee of humans to be published in a philosophy journal. Here is how this might work. Laura and Bill are logicians who believe there is an understudied problem in logic. They prompt an AI to survey and categorize a vast literature in an interesting and original way. They review the manuscript, check the citations, and do all the normal things they would do if they were reviewing their own or another human's paper. If they think it is excellent, they send it to a journal as an AI submission, vouching that the article is up to snuff. The editorial team would then review it as if it were a normal submission.[14] From there, the normal editorial process would ensue–reviewers would be consulted, decisions would be made, and so on. We are open to a variety of conventions regarding authorship that journals could use. One option would be to name the AI that produced the manuscript in addition to the human sponsor(s), e.g. Bard via Laura and Bill.[15] That said, we are not primarily concerned with authorship conventions in this paper.[16]

We think there should also be a limit set on the number of AI submissions that are reviewable. This could be accomplished in several ways: rules could be established whereby (i) journals limit themselves to only reviewing a certain number of AI submissions per year, (ii) philosophers can only sponsor one AI submission per year, (iii) submissions

[14] An anonymous reviewer worries that AI will have a hard time generating logic papers like these because it confidently makes major mistakes at basic points in proof that are hard to spot. While we believe AI will eventually develop to overcome basic issues like these, our argument does not rely on this prediction. Our argument is that *if* AI ever develops to produce sufficiently excellent papers, journals should publish them.

[15] How much work must one put into a paper to be considered a co-author? At this early stage in AI's development, it may be unclear at what point humans should be named as co-authors (as opposed to mere sponsors). Presently, AI needs prompting, and prompt engineering is a valuable skill, a skill that could give humans a claim to co-authorship rather than mere sponsorship. In the future, when AI requires less sophisticated prompting, human prompters will likely have a weaker claim to co-author status. For the time being, we are open to a variety of conventions regarding the co-authorship threshold. In any case, the vagueness of this threshold is not a unique problem for AI papers–the co-authorship threshold is vague in human-only papers as well.

[16] For a provocative new argument about why we should replace authorship with a bundle of roles that contribute to authorship, see Habgood-Coote 2024.

must be sponsored by three or more human sponsors, or (iv) if a human sponsors an AI paper to a given journal, then they cannot submit one of their own articles to that journal for a year. Many similar rules could be formulated, and each set will come with incentives and trade-offs. Editors will need to think carefully and creatively about how to structure such rules, and those rules may change over time depending on the state of academic publishing and academic philosophy. In any case, we strongly recommend a limiting rule so that journals are not inundated with AI submissions.

Whatever process journals adopt, it is important that humans remain in the process. We do not want a future where AI submissions are reviewed exclusively by AI reviewers. We worry that 'AI philosophy' could devolve (or evolve) into something that is too far removed from the projects human philosophers are engaged in or care about. It would be a shame to open a philosophy journal and find that AI left humans in the dust with its lingo, conceptual taxonomies, and points of focus.[17] This would undo the benefits that we were after in the first place. We want journals to publish excellent philosophical work because it helps humans philosophize. We want AI to aid us in this mission, so we must ensure their inclusion is compatible with that mission.

4. Objections

Remember, our thesis is that if AI can produce excellent philosophy papers, philosophy journals should strongly consider publishing those papers. Those who disagree will likely do so on one of two grounds: that the potential costs of AI publications outweigh the potential benefits or that AI submissions should, in principle, not be accepted. We will begin by addressing the former line of reasoning.

A. AI submissions will disincentivize human philosophy

When we first considered this issue, our primary concern was that AI submissions would disincentivize humans from doing philosophy. This would be a significant cost. The objection goes like this: If journals begin accepting AI submissions, AI may eventually crowd humans out of journals. If AI can produce one excellent article, it can likely produce many excellent articles. These submissions would quickly inundate

[17]Thanks to [redacted] for bringing this objection to our attention.

journals, making it nearly impossible for humans to publish.[18] If it becomes nearly impossible for humans to publish, humans will be disincentivized from doing philosophy. Philosophizing is essential to human flourishing, so AI submissions would result in a significant cost. We can formulate this argument as follows:

(1) Allowing AI submissions would likely crowd humans out of journals.
(2) If humans are crowded out of journals, then humans would be disincentivized from doing philosophy.
(3) Disincentivizing humans from doing philosophy is a significant cost.
(4) So, allowing AI submissions would likely result in a significant cost.

If journals accept the Sponsorship Model, this objection will not get off the ground. Journals could limit AI submissions by tying them to human sponsors to prevent a crowding-out problem. However, for the sake of argument, let us grant (1); let us consider the extreme scenario, a world in which AI submissions completely crowd humans out of journals. Let us also grant (3): we agree that disincentivizing humans from doing philosophy is a significant cost. That leaves (2). If humans are crowded out of journals, does it follow that humans are disincentivized from doing philosophy? We think not. In fact, AI submissions may even incentivize human philosophy.

There are many ways of doing philosophy: teaching, reading, writing, thinking, discussing, etc. Before any of us published academic articles, we took ourselves to be doing philosophy. Undergraduates do philosophy when they stay up late in their dorms debating the existence of God. Teachers do philosophy when they teach philosophy classes. We do philosophy when we write down our ideas with no eye toward publishing. We do philosophy when we help a friend wrestle with a deep question about morality or the political system. These ways of philosophizing are arguably more valuable than publishing in journals. So, in the (unlikely) world where AI truly crowds us out of publishing, human philosophers now have time on their hands to do philosophy in these ways. Efforts previously devoted to publishing could be diverted to teaching, reading, discussing, etc. Instead of devoting time to another niche publication for tenure, one could improve their classes, build relationships with students, or volunteer to teach philosophy in prisons. This will

[18] AI submissions could make it impossible for humans to publish because either (i) the sheer number of AI submissions destroys the journal system or (ii) the number of AI submissions crowd humans out while the journal system itself remains intact.

especially be the case if, as we argue in the next section, academic hiring practices adjust to this world, where administrators would deemphasize research output for hiring and promotion decisions.[19]

Additionally, most philosophers already face a crowding-out problem. Philosophical geniuses increasingly publish and crowd the rest of us out of philosophy journals. Yet, their work does not disincentivize the rest of us from doing philosophy. Yes, we are less likely to publish in such journals, but we also benefit from enjoying, discussing, teaching, and thinking about great philosophical work. If anything, the work of geniuses incentivizes human philosophy. A similar worry existed in chess; many thought AI would ruin the game for humans, but instead, it seemed to have the opposite effect (Keener 2022). So, even if humans are crowded out of journals – something we doubt will happen on the Sponsorship Model – we have good reason to think humans will not be disincentivized from doing philosophy.[20]

B. AI submissions will hurt the philosophy job market

We could reframe the previous objection as a concern about the academic job market. One might think AI submissions will disincentivize philosophy because AI submissions will likely put philosophers out of jobs. If AI crowds humans out of journals, humans would be unable to publish. Without journal publications, why would universities keep us around? We can formulate the argument like this:

(1) Allowing AI submissions would likely prevent humans from publishing in academic journals.

[19]One way to illustrate this point is that formal departmental expectations of research, service, and teaching usually define academic work. From rational self-interest, the way academics should prioritize these branches is in accordance with how much weight their departments give them in decisions about tenure and promotion. For example, a department that essentially guarantees tenure if an applicant has a handful of elite journal articles will incentivize junior faculty to prioritize research over teaching and service. But in an unlikely world where humans cannot publish, teaching and service are the only two formal departmental expectations that make sense. In this unlikely world, academics are incentivized to spend much more time on teaching.

[20]An anonymous reviewer observes that being disincentivized from philosophical *publishing* might cost us greatly. Our responses: (1) In the unlikely world where AI crowds out humans of academic journals by publishing much better work, we still think humans would want to write about and share their views about philosophy in other informal contexts. After all, many of us care deeply about philosophical writing for its own sake. Many philosophers write on blogs or have a Substack, and we see no reason why this would not continue. (2) We want to continue stressing that AI will likely not crowd humans out of academic journals. Even if AI can publish papers, humans will likely be long involved in prompting and co-authoring with them. Part of the reason for this is that humans will likely have an edge over AI in producing creative, beautiful, and engaging philosophical writing. (3) We reiterate the *philosophical geniuses* response: most of us are already competing with a small group of elite philosophers who occupy most of the space in elite journals. This fact does not obviously disincentivize everyone else from trying to publish, too.

(2) If humans are less likely to publish in journals, then the job market would suffer, i.e. fewer philosophers would hold academic posts, or philosophers who already hold academic posts would lose job security or competitive compensation.
(3) A suffering job market would be a significant cost.
(4) So, allowing AI submissions would likely result in a significant cost.

For the sake of argument, let us concede (1) the extreme scenario. Let us also grant (3): a suffering job market is a significant cost. It is a cost because if philosophers do not have jobs that allow them to philosophize, humans, in general, would probably do less philosophy. Once again, we are skeptical of (2).

We think (2) is a dubious prediction for several reasons. The proponent of (2) would need to make the case that (i) the job market would suffer more than it otherwise would, (ii) that this additional suffering would be due to the acceptance of AI submissions in philosophy journals, and (iii) that the job market would suffer in the *long term*. This latter point is the most difficult to prove. The case for short-term suffering is straightforward: current market practices rely on publication records as a heuristic for hiring and tenure decisions. If humans cannot publish, it will likely hurt their chances in these decisions. For example, if tenure requires six prestigious publications, but AI takes over the year before Maria is up for tenure, Maria may only have five publications. It is possible the tenure committee would deny her case, which would hurt her and her prospects for philosophizing. However, this argument only works until the market adjusts, and this adjustment seems inevitable. It seems likely that the tenure committee would eventually become aware of the AI takeover and adjust its standards. The publication heuristic would become useless for such decisions if no human had published in the past several years. How would a hiring committee decide between several candidates who have yet to publish in the past three years? Such committees would need to rely on other heuristics like a modified publication heuristic (e.g. publication record before AI publications) or records of excellence in teaching, service, or public philosophy. This market adjustment seems unavoidable, which means that, in the long term, there is little reason to think an AI publishing takeover would damage the job market as a result of hiring or tenure decisions. The market will likely correct to downplay research in exchange for teaching and service.[21]

[21] An anonymous reviewer observes that this will be bad for philosophers with research positions, especially those who are unskilled teachers. In some sense, we agree: the market may come to care

Though we think such a takeover is unlikely, we also think the market changes it would bring about would be good. First, consider this observation: 'Quite a bit of intellectual talent and energy is being channeled into producing thousands upon thousands of papers and books that hardly anyone will ever read or want to read'. (Huemer 2017) Much of this is driven by the frantic search for job security. Philosophers often lament the market incentive structure that drives pre-tenure philosophers to hyper-specialize and over-submit. Worse, this hyper-specialization results in work that only a few people read, which makes academic publishing feel like a strange and inconsequential enterprise. So, if market forces change hiring heuristics, many will welcome this development. They can divert their time toward the other forms of philosophizing we discussed in the previous section.

There is even a case to be made that an AI publishing takeover would bring about higher enrollment in the humanities, which would certainly benefit the job market. By the time AI can produce excellent philosophical work, it will likely be able to produce excellent work in various other disciplines. For example, the invention of the steamboat not only changed how boats were powered but also allowed us to use a variety of factory machinery that affected all kinds of different jobs (Brynjolfsson, Rock, and Syverson 2017; Danaher 2019, 43–44). The invention of sophisticated Large Language Models will have a similar effect. If they become sophisticated enough to outpace humans in their abilities to write excellent philosophical papers, they will likely change dynamics in other fields. If this happens, college students might be incentivized to pursue a major that markets skills that are less amenable to full automation. For example, the tech industry is progressively automating many jobs that were originally for humans. One result of this development is that an undergraduate degree in computer science can no longer virtually guarantee a job for its graduates (Korducki 2023). Those who want to work in the tech industry must either give up that aspiration or lean into other skills they have

about other skills in an academic philosopher beyond publishing, which will make philosophers who only know how to publish research less attractive. But we can soften this blow in several ways: first, this group is quite small; the majority of academic philosophers do not have research positions. So, this is not a problem that affects most philosophers. Second, many of these philosophers have the protection of tenure, which means they would likely only lose their position if they abjectly fail in their teaching responsibilities. Third, beyond stereotypes of researchers as terrible teachers, we do not have great evidence to think most in this group are bad teachers. In fact, there is good reason to think they are probably good teachers. We know that being a good researcher does not entail being a good teacher, and vice versa. But we also know that the skills required for producing excellent research also seem important for being a good teacher: good reading, clear communication, mastery of literature, etc. So, it's unclear if this is a serious problem even for philosophers with established research positions.

that AI still needs to develop. The humanities is one of the places in the academy that is the least amenable to automated labor, even though AI might automate certain features (e.g. some parts of academic publishing). This is likely why some entrepreneurs with a degree in philosophy have enjoyed so much success in non-philosophy industries (Gregoire 2014). AI can easily out-code these entrepreneurs, but the entrepreneurs possess skills in creativity, marketing, and problem-solving that even the most advanced AI models lack in spades.

But one might object: what if AI takes over teaching, too? What if AI publications are the gateway to AI philosophy teachers? To an extent, this is already happening in some disciplines (Shaw 2023). Even so, we contend that we should *not* allow AI to take over philosophy teaching. Notice that this claim is consistent with our thesis. We hold that AI work could be published in philosophy journals and that AI should not be permitted to teach philosophy classes. AI publications need not entail AI teaching. We think there is unique, irreplicable value in the human-to-human, student-teacher philosophical dialectic, whereas in the case of academic publishing, the goods associated therewith can be preserved or replicated in other ways.[22]

The market is hard to predict, especially following a major technological innovation. Our purpose in speculating is to show that a plausible case can be made that AI publications will ultimately help the job market, not hurt it. That said, our response here cedes an AI takeover of publishing, a takeover we think is highly unlikely. While AI is developing rapidly, we are doubtful it will develop fast enough to generate full-length papers worthy of publication anytime soon. If AI gets into publishing, we expect humans will be quite involved in its output for a while. After all, AI still has a deep problem with confidently publishing false information, a problem that does not look like it will be solved anytime soon (Hicks, Humphries, and Slater 2024: 3, Davis and Aaronson 2023). Moreover, journals do not just prize good arguments; they prize *excellent* philosophy, which often includes the creative and beautiful writing that AI will likely struggle to develop for a while. With these considerations, we still think humans will stay involved in the publishing game even when AI develops far beyond its capabilities now.[23] This is especially true if our Sponsorship

[22]But for a dissenting opinion, see Danaher 2023 and Danaher forthcoming for a defense of the possibility that AI is capable of interpersonal relationships.

[23]An anonymous referee objects that our proposal loses out on some important functions we attribute to authorship, such as 'creating an intellectual market' (Habgood-Coote 2024, 2.5). This function is 'assigning a set of people as the authors of a paper … to create a system of private goods which are apt for market mechanisms' (14). Authors seem to need incentives to write helpful content for

Model is the one most journals adopt. Nonetheless, we hope to have shown that our relationship with philosophy might improve even in the worst-case scenario. Of course, if the worst-case scenario never obtains, the objections from incentives and the job market are much less powerful.

C. AI cannot take responsibility for its work

Several publishers have recently claimed that AIs should not be named as authors because AIs cannot take responsibility for their work. For example, Magdalena Skipper, editor-in-chief of *Nature*, said, 'An attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs'. (Stokel-Walker 2023). Other philosophers have followed suit. Habgood-Coote et al. claim, for example, that one of the functions of authorship is to hold them accountable, 'to create a target for praise if the paper is epistemically good, and censure if the paper is epistemically bad' (2024, 13). Such a practice is beneficial because it incentivizes authors to make sure they produce good work (ibid). Consider, for example, Harvard's Francesca Gino, a business school professor recently accused of falsifying data in her research on dishonesty (Kim 2023). It seems important that we can point to *her* and call *her* to account for her actions. As Habgood-Coot et al. note, this is important because we can disincentivize Kim and other scholars from following suit (2024, 13). However, some might worry that our proposal misses this important function. If a journal publishes a terrible AI-generated paper, it seems there is nobody to hold accountable. So, one cost of our proposal is that we miss out on the accountability we need to incentivize quality control among authors. Without such accountability, we might have no

the profession, and this function helps meet that incentive. However, because AI does not need a similar incentive, this function does not apply to them. We have two responses: First, as we note, humans will likely stick around in publishing for a while anyway. The social credit we attribute to humans who either sponsor or co-author with AI might need some tempering, since, arguably, a solo-authored human paper deserves more reward than a co-authored paper or a sponsored AI-generated paper. But even this is not a sure thing: a sponsor's ability to skillfully prompt AI to generate a sufficiently excellent paper might be praiseworthy, such that skillful sponsors may deserve social credit, too. Second, all the objection shows is that some functions serve human authorship, and some do not transfer over to authorship attributed to AI. We do not see what is objectionable about this fact. The function is either inapplicable to an AI-generated paper or somewhat applicable to the human sponsor who prompted the paper. But if the objector wants to press that this is a necessary condition of authorship, we are also happy to demote AI to something less than an author – maybe a guarantor, a generator, or some alternative title (see ibid: section 3 for a discussion of alternatives). This is no problem. Journals can call AI whatever they would like; we are just proposing that those journals should publish their outputs.

way to control bad things in publishing, like plagiarism, terrible arguments, and the like.

We agree that accountability is an important feature of academic publishing, but we deny that our proposal leaves no room for it. After all, the editorial team is also responsible for the work they publish, which bears out in the real world. Several years ago, *Hypatia*, a feminist journal of philosophy, published a controversial article on transracialism (Tuvel 2017). Critics of the article claimed that it demeaned transgender people and Black people. The editorial team was widely criticized and eventually apologized (Jaschik 2017). When an editorial team publishes an inappropriate article, we hold them accountable. So, even in a world where AI is publishing papers, the editorial team serves as a party we can hold accountable for the contents of the papers it publishes. In an important sense, however, this does not quite match the function of accountability for authorship as Habgood-Coot et al. put it. In their view, accountability incentivizes the paper's author, not the editorial board that published it. Luckily, our Sponsorship Model requires that journals only consider AI-generated papers if sponsored by an accompanying human who can vouch for its contents. So, in a case where an AI-generated paper is inappropriate in some respect, we can divide the accountability between two parties: the editorial team and the human sponsor. One benefit of this division of accountability is that the onus is not entirely on the editorial board. This is especially important since we cannot reasonably expect reviewers and editors to be familiar with all the sources and facts AI could get wrong, make up, or even plagiarize. If we can also hold human sponsors accountable, then the division of labor in verifying the content of the AI-generated paper is much more manageable. In fact, because the human sponsor is the one who took the initiative to submit the paper, it seems reasonable to primarily hold the sponsor accountable, even if the editorial team still shares some responsibility for the contents of the papers they publish.

Moreover, other publishing practices reveal that author accountability is not strictly necessary for publishing. Consider, for example, anonymous, pseudonymous, and posthumous publications. Anonymous and pseudonymous publications are occasionally found in academic publications (e.g. *Journal of Controversial Ideas*).[24] Such authors cannot be held directly accountable for their published material, i.e. the paper's audience cannot

[24] *Journal of Controversial Ideas* confirmed via email they consider pseudonymous submissions where even their own editorial team is ignorant of the author's true identity.

find a target of praise or blame that would incentivize the author to do a better job, which is the purported function of accountability in authorship (Habgood-Coote 2024, 13).[25] The same is true of posthumous publications. We regularly publish the work of the deceased even though we can no longer hold them accountable. In fact, we would find it odd if journals *always* required author accountability. Consider this hypothetical: Smith submits an excellent article to a top-tier journal, and the article is accepted without conditions and is forthcoming. Smith tragically passes away in a trolley accident. The journal editors hear of Smith's accident. Should they rescind the planned publication of Smith's article? Of course not. Though Smith cannot be held accountable for his work, this is insufficient reason to rescind his work's acceptance to the journal. So, even if the Sponsorship Model does not solve the accountability problem, we have good reason to think accountability is not necessary for publishing anyway.

In sum, our proposal does not miss out on accountability as long as a human sponsor and an editorial team are attached to the publication of the AI-generated paper. In cases where an AI-generated paper is published with inappropriate content, the journal can do what it already does when human-written papers are published: either rescind the paper or allow the author to issue an erratum. Additionally, there are many cases where we publish work without the ability to hold the author accountable, so we see no reason to prohibit AI on these grounds.[26]

D. AI submissions would cause a significant loss in reviewers

[25]For an interesting discussion of pseudonymous philosophy in the history of philosophy – from the various historical "pseudo" authors (e.g. Psuedo-Dionysius) to analytic figures like David Lewis and Amelie Rorty – see Egid 2023.

[26]An anonymous reviewer raises a nearby objection: The status quo allows us to assume good faith with human authors – i.e. we can assume authors are not just trolling journals when they send in papers, that they are not just cracking jokes or throwing in blatant lies in random parts of the paper. But with AI, things are different. Unlike humans, we have no assumption of good faith from AI, which would make reviewing their papers quite arduous. Therefore, allowing the entrance of AI without a good faith assumption would make an already arduous peer-review process much more difficult than it already is. Our responses are as follows: First, this is a contingent problem that we believe Large Language Model creators are motivated to minimize. After all, the various corporations compete with one another in the market, and one of the ways they can get ahead of the others is by programming their models to produce output that is comparatively more accurate than the others. That is not to say the models will suddenly have good faith in virtue of having heightened capabilities, but it does at least mitigate the worry about just how much reviewers will have to fact-check the AI-generated papers. Second, we think reviewers are, in some sense, required to do this basic fact-checking. If an author gives an argument, part of the reviewer's job is to verify the argument for validity and soundness. If the reviewer fails to do this, we think they are being derelict in their duties as a reviewer. Third, on our Sponsorship Model, we can still assume in good faith that the sponsor has vetted the paper before she has submitted it. This is also not a total fix, but it does give us another way of softening the objection.

Another potential cost of our proposal is that it might cause a loss in reviewers. Some philosophers may lament AI's entrance into the humanities, and they might refuse on principle to review articles from journals that accept AI submissions. There may also be other philosophers who are less pessimistic about AI but nonetheless have reason to refuse to review for journals that are open to AI. After all, one of the main reasons any philosopher reviews any paper is to provide a service to other philosophers.[27] However, if we began reviewing AI-generated papers, we may either be doing no service at all for any philosopher (because the paper was entirely AI-generated) or only a small service for the human sponsor that prompted the paper. This would be a significant cost for philosophy journals, especially given that one of the most popular complaints about academic publishing is how hard it is to get reviewers.[28]

We have several ways to soften the blow of this objection. First, reviewing AI-generated papers is still a way for reviewers to serve other human philosophers. When journals publish an author's paper, that is not just a benefit to the author; more than this, the publication is a benefit to others who may either enjoy reading the paper or build on the argument of the paper to make progress on some important question. So, when a journal evaluates an excellent AI-generated paper, it evaluates a paper that can still massively help other human philosophers (not to mention the human sponsoring the submission). Second, this is not an in-principle objection to our view. We can imagine worlds where, for irrational reasons, reviewers refuse to review papers from left-handed authors, which would cause a massive referee shortage. This fact gives journals prudential reasons to avoid reviewing left-handed authors, but it does not give an in-principle reason that it would be morally wrong for them to review left-handed authors. In a world like this, our thesis is still true: if an AI-generated paper is sufficiently excellent, journals have a strong reason to publish it. Third, journals could solve this problem by keeping peer review anonymous and refusing to disclose that the paper is AI-

[27]We thank an anonymous reviewer for raising this objection.

[28]One additional way our proposal could overwhelm journals is by potentially allowing a near-infinite pool of submissions to flood submission queues. In a world where AI can generate a sufficiently excellent paper, they will likely be able to do this significantly faster than humans. The worry is that journals might get overwhelmed by the (already present) problem of massive submission pools. Luckily, the Sponsorship Model solves this as long as a human sponsor is required and journals retain the popular policy that they will only consider one submission from a human at a time. With these two requirements, there is a reasonable limit on just how large the submission pool can get. Of course, if journals do not use the Sponsorship Model or do not enforce their normal rule about submission limits, they open themselves up to this objection. We see this as all the more reason to endorse the Sponsorship Model.

generated to reviewers. After all, there are probably some cases where a philosopher would refuse to review someone's paper if the editorial team disclosed relevant features of the author's identity to the reviewer. However, we know such disclosure is inappropriate because the reviewer's job is to evaluate the paper on its merits, not decide whether the author has any features that militate against her paper's publication. The same dynamic should carry over to our proposal: reviewers evaluate the paper's arguments, not the author herself. Of course, if a journal's decision to refuse to disclose whether a submission is AI-generated causes referees to refuse any request altogether, then that is a strong reason for the journal to pause its policy. But here, we kick back to our second reply: this is not an in-principle objection to our view that journals have strong reasons to publish excellent AI-generated papers, but only a prudential objection.

E. AI submissions violate publishing norms

Now, let us address some in-principle objections to AI submissions. We have argued that philosophy journals generally have this primary goal: to publish excellent philosophy articles. Of course, rare cases exist where an *otherwise* excellent article should not be published because it violates some publishing norm. For example, articles that incite violence or explicitly demean others should not be published *even if* they are excellently written and uniquely contribute to some field. Many publishing norms are unspoken, subtle, and vary across academic circles. For example, one might think publishing norms require that an author be convinced beyond a reasonable doubt that his or her article's claims are true. Another might think publishing norms only require that an author believes his or her article moves the conversation forward. Others might think that no such norms exist: an author can publish an article that he or she thinks contains weak arguments as long as others deem the article worth publishing. There is a surprisingly large debate over this norm.[29] We will now investigate if other such norms would make our proposal undesirable.[30]

[29]For arguments against this norm, see Dang and Bright 2021 and Plaikas 2019. For arguments for this norm, see Sarıhan 2023; Buckwalter 2022, and Basu 2024.

[30]We might ask: 'How do we know which norms should govern the profession?' There are, broadly, two ways we can know: The first way is via *cost-benefit analysis*, where we make a prudential judgment about whether the benefits our proposal can realize will sufficiently outweigh whatever costs come along. We used this method in the first part of the paper, where we considered consequentialist objections. However, the second way is applying *in-principle* moral norms to our proposal (e.g. if it is

(i). The norm against bullshitters

Here is one way our proposal might violate a publishing norm: AI cannot do philosophy correctly because Large Language Models are Frankfurtian bullshitters unconcerned with the truth of their outputs. (e.g. Bergstrom and Ogbunu 2023, Hicks, Humphries, and Slater 2024). They are not merely hallucinating, lying, or confabulating when they produce a false output. Instead, they don't care about the truth at all. Academic journals, an objector might say, should only publish authors who are concerned with the truth of their utterances, not authors who are indifferent to the truth of the claims in their work. So, because journals should not publish bullshitters, they should not publish AI-generated papers. Let us get more specific about the sense in which Large Language Models might be bullshitters; after this, we can assess whether this objection would pose a problem for our proposal.

Hicks et al. offer a helpful clarification of the ways we can speak of how AI might bullshit. The general, broad definition of bullshit is 'Any utterance produced where a speaker has indifference towards the truth of the utterance' (2024: 5). There are two more specific ways someone can count as bullshitting: *soft bullshit* is the sort produced without the intention to mislead the audience about the utterer's agenda, whereas *hard bullshit* is the sort produced with such an intention (ibid). One example of soft bullshit is the overconfident undergraduate who dominates a discussion about a text despite reading it very poorly. Here, she is not trying to deceive the class about her purposes; as far as she knows, she is just speaking about what she read in the text. Nonetheless, she is not very concerned with whether her testimony about the text is true; otherwise, she would have read it more carefully. Instead, she may have motivations that she is unaware of: she wants to look impressive, she talks a lot when she is anxious, or something else. But *hard bullshitters* do something extra: they are not only unconcerned with the truth of their utterances, but they are also trying to deceive their audience about this fact. Used car salesmen who talk up cars to desperate customers just so they can score a profit are hard bullshitters because they want their customers to believe that they genuinely like the car despite having no real beliefs

dishonest, it might be a reason to reject it even if it produces good outcomes). These are not mutually exclusive methods, especially since deontologists who are not absolutists will say some moral norms can give way if the outcomes are severe enough. We do not pretend to offer an uncontroversial and ecumenical way of figuring out what to do if a norm violates a general moral principle or if it gives us great benefits. Still, we hope to offer an intuitive case that proves our proposal is compatible with both general moral principles and the production of great benefits.

about it. With this distinction, Hicks et al. claim that Large Language models are definitely soft bullshitters and are probably hard bullshitters, too (ibid: 6–8). After all, these models have no concerns at all, which means they are unconcerned with the truth of their utterances and are thereby soft bullshitters. But they probably count as hard bullshitters as well because they are designed to give their audiences the impression that they are human-like agents trying to speak the truth, when in reality, they are just designed to give the *impression* that they are such agents (ibid: 2).

Our initial reply is that we agree AI counts as a soft bullshitter, but we are reticent to accept that it counts as a hard bullshitter. As the authors concede, such a position 'requires one to take a stance on whether or not LLMs can be agents', and we are very skeptical that they are (ibid: 5). AI can fail to count as an agent and still produce soft bullshit insofar as it produces an utterance with no concern for the truth of that utterance. However, the additional positive requirement that AI is also attempting to deceive its audience about the nature of its enterprise is a much more controversial thing to attribute. So, we want to note that one cost of this view is that it does require a controversial attribution of agency to Large Language Models at the expense of other alternative interpretations of these models (e.g. their false utterances are 'confabulations', per Edwards 2023).

Nonetheless, let us grant that Large Language Models are bullshitters in both senses. Even with this concession, we still deny that this is a problem for our view. We do not presently see any publishing norm against bullshitters. Thus, instituting such a norm concerning AI would be *ad hoc*. Provided an article is excellent, journals do not verify authors' intentions. After all, many authors publish papers for many reasons that have nothing to do with communicating true testimony. Some publish to secure tenure, look fashionable among their peers, secure job prospects, or even shit-stir (Agar 2021). Many philosophers publish work they are unsure about or have changed their minds about. Editors do not inquire about the authors' motivations to publish a paper; instead, they ask themselves whether the paper is excellent. If that paper is excellent, we think most editors count that as a decisive reason for publication. To see this, consider another case:

> Gettier Case: Suppose we found out that Edmund Gettier's 'Is Justified True Belief Knowledge?' was bullshit. He only wrote the paper to get tenure and was so indifferent to its success that he didn't even want to attend a conference

> celebrating its publication (Pritchard 2018, 23). Gettier was genuinely unconcerned about whether the paper's main conclusions were true, but wrote as if he was because he knew this raised his chances of publication.

In reality, Gettier was almost certainly concerned with the truth of his original paper. But in this stylized example, we imagine he was unconcerned. Arguably, this was dishonest for Gettier to do.[31] But our question is not whether bullshitting is honest or dishonest; instead, our question is whether the fact the paper is the output of a bullshitter means publishers should not accept the paper. We don't think so. In this case, the paper was excellent and revolutionized contemporary epistemology. Failing to publish the paper would deprive the field of immense progress on the question of epistemic justification, even if such progress was the consequence of published bullshit. Pair this with the fact that editorial teams do not inquire about authorial motivation before they publish, and we get a strong case that there is not a norm against bullshitting in publishing.

(ii) . The norm against platforming notoriously unreliable authors

Here is another objection near the previous one: Journals should not platform notoriously unreliable authors. Consider this case:

> Fake News: Marc Morano is a high-profile climate change skeptic who works for the Committee for a Constructive Tomorrow, an organization that promotes climate change denial. Despite having no scientific training, Morano is regularly interviewed on US news programs to provide balance to the expert opinion on climate scientists (Rietdijk and Archer 2021: 1). However, in one of his recent interviews, Morano fortuitously gives an *excellent* report about the COVID-19 vaccine.[32]

Morano is generally an unreliable testifier; he regularly says things publicly that are widely dismissed by the scientific community. While news networks continue to host him to represent an opposing side to the widespread consensus among scientists about climate change, there is a plausible case that networks should stop hosting Morano. After all, if a network continues hosting an unreliable testifier like Morano, the audience will see the network as no longer trustworthy or believe things they ought not to.[33] This seems the case even if Morano gets lucky in

[31] See Miller 2021, 19–20 for a brief discussion about bullshitting and honesty.
[32] This last detail is an embellishment; we do not know if Morano actually gave an excellent interview about the vaccine.
[33] Thanks to an anonymous reviewer for raising this objection.

an *excellent* one-off interview about another scientific issue. While Morano got things right in this one unique case, his general unreliability is a strong reason why networks should not continue hosting him.

Perhaps there is a parallel case with the publication of Large Language Models in academic journals. There are plenty of examples of AI getting things wrong, even producing so-called 'Fake news'. For example, a news reporter recently asked ChatGPT to generate an article that covers Michael Bloomberg's post-election activities (Glorioso 2023). While ChatGPT accurately recounted Bloomberg's work, its output was also littered with fake quotes from Bloomberg and his fictional critics. Given this case, ChatGPT seems like an unreliable author who spreads misinformation like Morano. Even if ChatGPT gets lucky and produces an excellent article for journal publication, its general unreliability is a strong reason why journals should not publish it.

While we feel the pull of this objection, it does not defeat our proposal. First, the relationship between a news outlet and its audience seems importantly different from the one between an academic philosophy journal and its audience. In the former dynamic, the audience generally trusts the news outlet's testimony about various events in the world. The dynamic is akin to the one between a theoretical authority on some complex issue and the audience that judges the authority as competent enough to merit their trust. (Of course, skeptical audience members may question a reporter's particular claim and independently verify it with other news outlets.) But in the latter dynamic, things seem different. Especially in *philosophy* journals, the proof of the author's main claims is supposed to be within the article itself, and the author is setting out to convince her audience to believe something (as opposed to merely reporting something). If an article lacks this proof, this is widely recognized among readers as a strike against the quality of the article. However, if the article contains the proof, the reader will usually have enough information to independently judge whether the article's main argument is sound. So, even if news outlets have reason to exclude unreliable authors like Morano, the dynamic between academic journals and their audiences leaves an open question about whether editors have a similar reason to prohibit unreliable authors.

Second, we are unaware of a publishing norm against unreliable authors in academic journals. Consider this case:

> Uncareful Thinker: Bill is widely known in the academic philosophy community as not being a very good philosopher. Everyone knows he gives terrible and

> incoherent talks at conferences, and the few papers he has published are known as some of the worst philosophy in print. But, to everyone's surprise, a top journal accepted Bill's most recent paper.

Bill has a well-earned reputation as an unreliable philosopher, yet that seems irrelevant to whether his paper is acceptable or not. In fact, an editor who is aware of Bill's reputation might even be delighted that Bill has finally written an excellent paper. If this is all right, we think things are no different for AI authors. Even if their terrible reputations precede them, their output merits publication if it is excellent.

Third, even if our previous responses do not work, the severity of this objection will substantially decrease as Large Language Models progress in their development. Right now, seeing ChatGPT mess up a basic *modus ponens* inference is unsurprising; the model has not been around very long and is still in the early stages of development. However, with several decades of development, this same mistake would be very surprising. After all, as we mentioned in the introduction, AI is already outpacing humans in various domains, even adequately mimicking philosophers such that experts have difficulty distinguishing between AI and a human philosopher (Schwitzgebel, Schwitzgebel, and Strasser 2023). And if we think about the skills involved in publishing certain types of papers (e.g. a straightforward reply paper that challenges a premise in another author's argument or a meta-analysis that summarizes what others have said on a certain topic), then it is conceivable that something like ChatGPT could join the publishing fold sooner than we think.[34]

5. Conclusion

To conclude, we have argued that if AI can generate an excellent philosophy paper, then philosophy journals ought to strongly consider publishing it. Remember, for a paper to be excellent, it must be original (not plagiarized). We have sought to move past questions of AI's theoretical capabilities to ask deeper questions about academic publishing and the purpose of philosophy journals. The benefits of excellent AI articles include the same ones that would come with excellent human articles. Moreover, AI may have unique abilities to survey vast swathes of literature

[34]The reputation of AI models as generally more or less reliable than their competitors is another way to shore up authorship's 'credibility judgment' function, per Habgood-Coote (2024). In their view, one function of authorship 'is to enable readers to make judgments about how credible the results of the paper are' (12). Readers can make these judgments by considering which model is associated with an AI-generated paper and considering the model's reputation comparatively to other models.

and identify problems and solutions that humans could not have identified. Of course, some costs come with significant technological change, but we proposed the Sponsorship Model to address some of them. Ultimately, we want humans to flourish. We think humans flourish when they engage in philosophy, and we think allowing excellent AI articles into philosophy journals is more likely to promote this end than to detract from it.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Agar, Nicholas. 2021. *Confessions of a Philosophical sh*t-Stirrer.* ABC Religion and Ethics.

Ayers, J. W., A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, et al. 2023. "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum." *JAMA Internal Medicine* 183 (6): 589–596. https://doi.org/10.1001/jamainternmed.2023.1838.

Basu, Rima. 2024. "Against Publishing Without Belief: Fake News, Misinformation, and Perverse Publishing Incentives." In *Attitude in Philosophy*, edited by Sandy Goldberg and Mark Walker. Oxford University Press.

Bergstrom, C. T., and C. B. Ogbunu. 2023. "ChatGPT isn't hallucinating, it's bullsh*t-ting." *Undark Magazine*. April 6.

Brynjolfsson, E., D. Rock, and C. Syverson. 2017. "Artificial Intelligence and the Modern Productivity Paradox." A Clash of Expectations and Statistics (NBER Working Paper 24001).

Buckwalter, W. 2022. "The Belief Norm of Academic Publishing." *Ergo: An Open Access Journal of Philosophy* 9.

Clay, Graham, and Caleb Ontiveros. 2023. "Philosophers Ought to Develop, Theorize About, and Use Philosophically Relevant AI." *Metaphilosophy* 54 (4): 463–479. https://doi-org.ezproxy.baylor.edu/10.1111/meta.12647.

Danaher, John. 2019. *Automation and Utopia: Human Flourishing in an Age Without Work*. Cambridge, MA: Harvard University Press.

Danaher, John. 2023. "Moral Uncertainty and Our Relationships with Unknown Minds." *Cambridge Quarterly of Healthcare Ethics* 32 (4): 482–495. https://doi.org/10.1017/S0963180123000191

Danaher, J. forthcoming. "The Philosophical Case for Robot Friendship." *Journal of Posthuman Studies*.

Dang, H., and L. K. Bright. 2021. "Scientific Conclusions Need Not Be Accurate, Justified, or Believed by Their Authors." *Synthese* 199 (3-4): 8187–8203. https://doi.org/10.1007/s11229-021-03158-9

Davis, E., and S. Aaronson. 2023. Testing GPT-4 with Wolfram Alpha and Code Interpreter Plug-ins on Math and Science Problems. arXiv preprint arXiv:2308.05713.

Egid, Jonathan. 2023. “From the Pseudo to the Forger: The Value of Faked Philosophy.” *Aeon*.

Glorioso, C. 2023, February 23. Fake News? ChatGPT Has a Knack for Making Up Phony Anonymous Sources. NBC New York. Retrieved from https://www.nbcnewyork.com/investigations/fake-news-chatgpt-has-a-knack-for-making-up-phony-anonymous-sources/4120307/.

Gregoire C. 2014 “Why Philosophy Majors Rule.” *HuffPost*. March 6. https://www.huffpost.com/entry/why-philosophy-majors-rule_n_4891404.

Guo, B., X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu. 2023. How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. https://arxiv.org/abs/2305.03195. Accessed May 14, 2023.

Habgood-Coote, Joshua. 2024. “What’s the Point of Authors?” *British Journal for the Philosophy of Science*.

Harris, M. 2015. “Mathematicians of the Future." *Slate Magazine*. March 15. https://slate.com/technology/2015/03/computers-proving-mathematical-theorems-how-artificial-intelligence-could-change-math.html.

Hicks, M. T., J. Humphries, and J. Slater. 2024. “ChatGPT is Bullshit.” *Ethics and Information Technology* 26 (2): 38.

Huemer, M. 2017. “On Philosophy’s Uselessness to Society." Daily Nous. March 8. https://dailynous.com/2017/03/08/philosophys-uselessness-society/

Jaschik, S. 2017. *Journal Apologizes for Article on “Transracialism.”* Inside Higher Ed. May 1. https://www.insidehighered.com/quicktakes/2017/05/02/journal-apologizes-article-transracialism

Keener, G. 2022. “Chess is Booming." *The New York Times*. June 17. https://www.nytimes.com/2022/06/17/crosswords/chess/chess-is-booming.html.

Kim, J. 2023. “Harvard professor who studies dishonesty is accused of falsifying data." *NPR*. June 26. https://www.npr.org/2023/06/26/1184289296/harvard-professor-dishonesty-francesca-gino.

Koebler, J. 2017. *The Rise of the Robolawyer*. *The Atlantic*. April.

Korducki, K. 2023. “So Much for ‘Learn to Code’: In the Age of AI, Computer Science Is No Longer the Safe Major.” *The Atlantic*. Accessed September 26, 2023. https://www.theatlantic.com/technology/archive/2023/09/computer-science-degree-value-generative-ai-age/675452/.

Lee, J., T. Le, J. Chen, and D. Lee. 2023. “Do Language Models Plagiarize?” In *Proceedings of the ACM Web Conference 2023 (WWW ‘23)*, 3637–3647. New York, NY, USA: Association for Computing Machinery.

Lysandrou, G., R. E. Owen, K. Mursec, G. L. Brun, and E. A. Fairley. 2023. Comparative Analysis of Drug-GPT and ChatGPT LLMs for Healthcare Insights: Evaluating Accuracy and Relevance in Patient and HCP Contexts. arXiv preprint arXiv:2307.16850.

Martin, L., N. Whitehouse, S. Yiu, L. Catterson, and R. Perera. 2024. Better Call GPT, Comparing Large Language Models Against Lawyers. *arXiv preprint arXiv:2401.16212*.

Miller, Christian B. 2021. *Honesty: The Philosophy and Psychology of a Neglected Virtue*. New York: Oxford University Press.

Pritchard, D. 2018. *What is This Thing Called Knowledge?* Routledge, 23.

Rietdijk, Natascha, and Alfred Archer. 2021. "Post-Truth, False Balance and Virtuous Gatekeeping." In *Virtues, Democracy, and Online Media: Ethical and Epistemic Issues*, edited by Maria Silvia Vaccarezza and Nancy Snow. Routledge.

Sarıhan, Işık (2023). Problems with Publishing Philosophical Claims We Don't Believe. Episteme; Rivista Critica Di Storia Delle Scienze Mediche E Biologiche 20 (2):449-458. https://doi.org/10.1017/epi.2021.56

Schwitzgebel, E., D. Schwitzgebel, and A. Strasser. 2023. "Creating a Large Language Model of a Philosopher." *Mind & Language* 39 (2): 237–259. https://doi-org.ezproxy.baylor.edu/10.1111/mila.12466.

Shaw, J. 2023. "Embracing AI : The dawn of the virtual teaching fellow." *Harvard Magazine*. August 10. https://www.harvardmagazine.com/node/84551.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, et al. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–489.

Sparkes, A., W. Aubrey, E. Byrne, A. Clare, M. N. Khan, M. Liakata, M. Markham, et al. 2010. "Towards Robot Scientists for Autonomous Scientific Discovery." *Automated Experimentation Journal* 2: 1–11.

Stokel-Walker, C. 2023. "ChatGPT Listed as Author on Research Papers: Many Scientists Disapprove." *Nature 613* (7945): 620–621. https://doi.org/10.1038/d41586-023-00107-z.

Tuvel, Rebecca. 2017. "In Defense of Transracialism." *Hypatia* 32 (2): 263–278. https://doi.org/10.1111/hypa.12327

Webb, Maria. 2023. "Magnus Carlsen: How Intuition and AI Shape the Best Chess Player in the World." *Techopedia*. August 2.

Wellman, M. P. and U. Rajan. 2017. "Ethical Issues for Autonomous Trading Agents." *Minds and Machines* 27 (4): 609–624. https://doi.org/10.1007/s11023-017-9419-4

Williams, K., E. Bilsland, A. Sparkes, W. Aubrey, M. Young, L. N. Soldatova, K. De Grave, et al. 2015. "Cheaper Faster Drug Development Validated by the Repositioning of Drugs Against Neglected Tropical Diseases." *Journal of the Royal Society: Interface* 12 (104): 20141289. https://doi.org/10.1098/rsif.2014.1289.