

Between Human Extinction and the Extinction of Good Arguments: Placing Warning Signs for the Survival of Both

Murilo M. Vilaça

Volume 9, Number 2, 2026

URI: <https://id.erudit.org/iderudit/1124217ar>
DOI: <https://doi.org/10.7202/1124217ar>

[See table of contents](#)

Publisher(s)

Programmes de bioéthique, École de santé publique de l'Université de Montréal

ISSN

2561-4665 (digital)

[Explore this journal](#)

Cite this document

Vilaça, M. M. (2026). Between Human Extinction and the Extinction of Good Arguments: Placing Warning Signs for the Survival of Both. *Canadian Journal of Bioethics / Revue canadienne de bioéthique*, 9(2), 165–167.
<https://doi.org/10.7202/1124217ar>

Article abstract

This text is a response to Torres' article on artificial superintelligence and human extinction. I place some warning signs on the author's argumentative trajectory, arguing that the topic of human extinction is relevant, so it is necessary to correct the course at several points.

© Murilo M. Vilaça, 2026



This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

<https://apropos.erudit.org/en/users/policy-on-use/>

érudit

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

<https://www.erudit.org/en/>

RÉPONSE À - ARTICLE / RESPONSE TO - ARTICLE

Between Human Extinction and the Extinction of Good Arguments: Placing Warning Signs for the Survival of Both

Murilo M. Vilaça^a

Texte discuté/Text discussed: Torres ÉP. [If artificial superintelligence were to cause our extinction, would that be so bad?](#) *Can J Bioeth/Rev Can Bioeth*. 2025;8(3):74-85.

Résumé

Ce texte est une réponse à l'article de Torres sur la super-intelligence artificielle et l'extinction de l'humanité. J'émetts quelques signaux d'alerte sur la trajectoire argumentative de l'auteur, affirmant que le sujet de l'extinction de l'humanité est pertinent et qu'il est donc nécessaire de rectifier le cap à plusieurs reprises.

Mots-clés

philosophie, bioéthique, revue de la littérature, faisceau d'idées fausses, bons arguments

Abstract

This text is a response to Torres' article on artificial superintelligence and human extinction. I place some warning signs on the author's argumentative trajectory, arguing that the topic of human extinction is relevant, so it is necessary to correct the course at several points.

Keywords

philosophy, bioethics, literature review, bundle of fallacies, good arguments

Affiliations

^a Department of Drug Policy and Pharmaceutical Assistance (NAF), National School of Public Health (ENSP), Oswaldo Cruz Foundation (Fiocruz), Rio de Janeiro, RJ, Brazil

Correspondance / Correspondence: Murilo M. Vilaça, murilo.vilaca@fiocruz.br

In his article, "If artificial superintelligence were to cause our extinction, would that be so bad?" (1), Torres focuses on the question of human extinction through the risk posed by artificial superintelligence (ASI). According to him, "this is a topic that [...] bioethicists have not adequately examined" and "philosophers lack a robust theoretical framework for providing nuanced answers to this question" (p.74), Torres presents his aim as modest: "to encourage more vigorous debate about this topic among bioethicists, and to do this by applying the theoretical framework that I have developed elsewhere to the particular case of ASI" (p.74-5).

It may sound very promising when someone emerges who can fill important gaps in knowledge on a complex and relevant topic, in two areas — bioethics and philosophy — that no one has been able to fill before. This certainly catches the reader's attention, since, searching for ("artificial superintelligence" AND "theoretical framework") and ("artificial superintelligence" AND ethics AND "theoretical framework") on Google Scholar, we have the following results respectively: 259 and 174 (from 2000 to 2024, in the search carried out on December 4, 2025).

Here, then, it is appropriate to place a first "warning sign". Torres's assertions about the inadequacy of the approaches of bioethicists and philosophers create a burden of proof for him, namely to demonstrate that in all the relevant literature, nothing can be considered adequately examined, nor is there a robust theoretical framework. Insofar as he appears not to have adequately reviewed the literature, nor present the reasons for not reviewing certain texts, nor identified the authors of the texts by area of expertise (bioethics or philosophy), Torres may have committed a methodological error or, ultimately, may have *petitio principii*, since he begins the text with assertions (which may be the core of his argument) that he does not properly demonstrate. While we can assume that not all the results of the above searches are relevant to answering the author's question, the reader is not presented with anything to help them understand to what extent the texts not reviewed by the author would not change a scenario of the debate that he presents as incomplete. Instead of proving that there is no good answer to the question, he states that it doesn't exist and that he will provide one. Put another way, on what basis does the author make the above claims about the limitations of bioethicists and philosophers, in such broad terms?

This warning sign is a big one. In several paragraphs, Torres uses declarative sentences, offering few or no references (sometimes just one; sometimes that single reference is to himself). This cannot be called a literature review. Despite that, on p. 75, for example, he announces the "consensus view" (which, in footnote 4, he informs us he has already called the "default view"). We are on the second page, and we are already faced with a supposed consensus (which was a default) on this controversial topic. The author doesn't explain why the terminology changed. "Now [I] prefer the [new] term", he says (p.75). Personal preferences are not exactly a good way to achieve the conceptual robustness the author claimed to seek.

Based on a fairly limited number of texts, he writes sentences, creating a highly questionable overall effect. "Transhumanists [...] would say that [...]" (p.78), "the vast majority of pro-extinctionists accept [...]" (p.78), and so on. Torres's transhumanists are reduced to two texts by Nick Bostrom, which, for any expert on the subject, is not even remotely convincing in presenting the perspective of this transhumanist alone.

According to Pirie, “the fallacy of composition occurs when it is claimed that what is true for individual members of a class is also true for the class considered as a unit” (2, p.31). Torres puts everything in the same “bundle” (3), as he states in footnote 12: “It is for this reason that one might wish to classify versions of transhumanism, longtermism, and other TESCREAL (Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalist, Effective Altruism, and Longtermism) ideologies as ‘pro-extinctionist.’” (1, p.80). Although he may claim that the “simple repetition of a point of view does nothing by way of supplying additional evidence or support” (2, p.111), Torres’s bundle becomes increasingly full of different things, and as a result, the conceptual gaps widen.

A second warning sign refers to Torres’s ASI and what I will call the epistemic gap. “Imagine” (1, p.75), the author proposes. Imagining is a wonderful human capacity. But, in an “ocean of possible imaginations” about what an ASI could be, a concept that has been used in very different ways, Torres seems to incur the lack of conceptual (and, I add, epistemic) robustness that he points out. Further on, in the “Equivalence Views” section, Torres ‘imagines’ two other extinction scenarios involving the ASI. It would be more coherent or prudent to write “involving an ASI” because, as he claims, “the details of Going Extinct are paramount” (1, p.79).

But there is another side to the epistemic gap that should be highlighted. In a recent publication, Agar and Vilaça (4) argued that human extinction via an ASI may be a “problem of charismatic extinction threats” and counsel the adoption of a “imagination insurance for an intrinsically uncertain future”. “We are, of course, free to imagine an ASI with such power [like Skynet]” (4, p.7), they note, but claim that “Imagination insurance offers a way to bring attention to extinction threats that overly focusing on [ASI] and other exotic extinction threats causes us to overlook. [...] Humanity needs imagination insurance in respect of extinction threats” (4, p.10-1).

Agar and Vilaça start from a fact to offer a relevant normative answer to a central question of our time: “how much of our finite pool of worry we should allocate to each extinction threat [...]. Only then can we decide which threats to ignore and which to seriously prepare for” (4, p.11). The authors’ suggestion isn’t exactly groundbreaking, but given the current scenario, it proves to be very useful: We should focus on the intrinsic causes of a potential threat of extinction, that is, in [new] evidence. Without providing or addressing evidence, our imagination can take us far from what matters.

Although he promises to offer a solution to the problem of “lack a robust theoretical framework”, Torres surprisingly fails to consider that he may have been captured by the charisma of a threat that may be far from relevant to thinking robustly about the most fundamental part of his question: Would it be good if everyone on Earth (read: human beings) were killed? In this regard, Torres’s conclusion doesn’t sound so innovative:

My hope is that this provides a helpful degree of clarity to a deceptively complex issue: nearly everyone — including most proextinctionists — would concur that the mass murder of everyone on Earth would be extremely bad (1, p.83).

For someone deeply interested in the topic of human extinction and concerned about the quality of the debate, Torres does not seem to have singled out the most threatened danger, nor to have offered major contributions to what he calls “consensus view”.

I conclude this commentary by highlighting a controversial characteristic of Torres’s argumentation in some of his texts (here is the third warning sign): labelling as eugenics what he wants to criticize. He uses this term frequently in his arguments. In another text, alongside Geburu, eugenic ideology was the very strong link between very different things: racism, xenophobia, classism, ableism, sexism, transhumanism, Extropianism, singularitarianism, (modern) cosmism, Rationalism, Effective Altruism, and longtermism (3).

In his text (1), in footnote 14, Torres uses the word, qualifying it in an innovative way: digital eugenics (1, p.80). Eugenics appears in footnote number 14, in which he states: “For a discussion about what our artificial descendants might be like, and the ethics of creating artificial descendants, see (49). Note that I object to the sort of “digital eugenics” [...] – that this paper explores.” (1, p.80). In another text, Torres states that “digital eugenicists want to do away with biology altogether” (5). The reference “49” in Torres’ article is a text by Lavazza and Vilaça (6). I invite the reader to read that article, in which, broadly speaking, the authors argue that in a hypothetical scenario of imminent and irreversible human extinction (the ultimate threat), it would be better to consider how to preserve some of the value of human beings by generating non-organic successors (based on silicon) than to do nothing. At no point in the text do Lavazza and Vilaça argue that we should do away with biology altogether, nor that this would be desirable or better, which eliminates any possibility of identifying the authors as digital eugenicists.

Torres uses the expression “artificial descendants” and correlates it with “digital eugenics,” stating that “I object to the sort of ‘digital eugenics.’” (1, p.80). In other words, Torres associates Lavazza and Vilaça with advocates of something like digital eugenics. But a clever reader will easily see that this is not true. Lavazza and Vilaça neither explore, much less defend, the creation of artificial descendants, postulating to do away with biology altogether. Therefore, the definition of “digital eugenicists” used by Torres himself does not apply to them. The authors are very clear: they are not advocating the extinction of humanity, nor its replacement by artificial descendants (who would be superior). Starting from the hypothetical premise “[...] that there is

an impending threat and that there is a will to address the issue of our potential extinction” (6, p.3), they propose that we consider what entity could succeed us, carrying as much of our human value as possible. The scenario clearly outlined by the authors does not allow any inference that their arguments are pro-extinction.

Transhumanists, pro-extinctionists, or anyone Torres wants to oppose sound reprehensible at the origin, when associated with eugenics, without even needing to explain in depth what this means in its various forms and contexts (2). Considering that labelling someone or some argument as eugenicist does not tend to legitimize them/it, it is important to pay attention to this strategy of the author, which can be framed in a myriad of fallacious ways.

In Torres’s article, there are some interesting clues about the problems of the debate surrounding the question of human extinction, such as the different conceptions of what it means to be human. The overview presented (three main positions within existential ethics) may also be useful for introducing the debate, but the criteria for defining why these are the main ones are unclear. Perhaps the most important thing for advancing this debate is to address controversial issues in the most analytical way possible, and the least ideological or prejudiced. Instead of framing authors and perspectives within a preconceived framework, labelling them/it, we must identify how concepts, arguments, empirical evidence, and imagination can be combined to generate the best possible understanding of the topic.

Thinking about the threats to the presence of humans (and nonhumans) on Earth is crucial today. We need to unite the billions of minds around this (4). Imagination is precious, criticism is necessary, but we will not save humans at the expense of good arguments.

Reçu/Received: 02/08/2025

Remerciements

Je remercie le National de Desenvolvimento Científico e Tecnológico (CNPq) pour son soutien à la recherche: APQ/PRÓ-HUMANIDADES (421523/2022-0); Chamada Universal (421419/2023-7); Bolsista de Produtividade em Pesquisa - PQ2 (315804/2023-8).

Conflicts d'intérêts

Aucun à déclarer

Publié/Published: 16/03/2026

Acknowledgements

I thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the research support: APQ/PRÓ-HUMANIDADES (421523/2022-0); Chamada Universal (421419/2023-7); Bolsista de Produtividade em Pesquisa - PQ2 (315804/2023-8).

Conflicts of Interest

None to declare

Édition/Editors: Aliya Affdal

Les éditeurs suivent les recommandations et les procédures décrites dans le [Core Practices](#) de COPE. Plus précisément, ils travaillent pour s'assurer des plus hautes normes éthiques de la publication, y compris l'identification et la gestion des conflits d'intérêts (pour les éditeurs et pour les auteurs), la juste évaluation des manuscrits et la publication de manuscrits qui répondent aux normes d'excellence de la revue.

The editors follow the recommendations and procedures outlined in the COPE [Core Practices](#). Specifically, the editors will work to ensure the highest ethical standards of publication, including: the identification and management of conflicts of interest (for editors and for authors), the fair evaluation of manuscripts, and the publication of manuscripts that meet the journal's standards of excellence.

REFERENCES

1. Torres ÉP. [If artificial superintelligence were to cause our extinction, would that be so bad?](#) Canadian Journal of Bioethics/Revue canadienne de bioéthique. 2025;8(3):74-85.
2. Pirie M. How to Win Every Argument: The Use and Abuse of Logic. New York: Continuum; 2007.
3. Gebru T, Torres, ÉP. [The TESCREAL bundle: eugenics and the promise of utopia through artificial general intelligence](#). First Monday. 2024;29(4).
4. Agar N, Vilaça MM. [On artificial superintelligence and the problem of charismatic extinction threats](#). Journal of Ethics and Emerging Technologies. 2025;35(2):1-12.
5. Torres ÉP. [Digital eugenics and the extinction of humanity](#). Tech Policy Press. 11 Jul 2025.
6. Lavazza A, Vilaça MM. [Human extinction and AI: what we can learn from the ultimate threat](#). Philosophy & Technology. 2024;37:16.