Philosophy of AI
Vol. 1, 2025, 59-70
10.18716/ojs/phai/2025.2338

ARTIFICIAL INTELLIGENCE, PERSONAL ONTOLOGY, AND EXISTENTIAL RISKS

Andrea Sauchelli[1 2]

[1] Lingnan University, Department of Philosophy, Hong Kong, HK,
[2] Lingnan University, Hong Kong Catastrophic Risk Centre, Hong Kong, HK

Abstract Some forms of Artificial Intelligence (AI) may kill us all, or at least irremediably maim our hopes to enjoy the benefits of significant technological advances in the future – or so some philosophers working on existential risk to humanity have recently claimed. But what do 'us' or 'our' stand for in this context – and why could the extension of these terms not include, say, (superintelligent) AI systems as well? I explore several foundational issues in the recent debates on existential risks related to AI, with a particular focus on the conceptual connections between ontological theories of our nature – answers to the question, 'What am I?' – and recent formulations of the notion of an existential risk for humanity.

Keywords
Artificial Intelligence; Existential Risk; Humanity; Personal Ontology.

1. Introduction

Many have raised worries that some AI systems, including those of a superintelligent variety (SAI), may pose catastrophic or even existential threats *to us* (Bales et al., 2024; Bostrom, 2012, 2013, 2014; Cappelen et al., 2025; Dung, 2024; Müller, 2020; Ord, 2020, pp. 138–152; Russell, 2019; Turchin & Denkenberger, 2020; Vold & Harris, 2021; Yudkowsky, 2008). One contribution of this paper to this debate is the idea that which ontological theory of our nature we adopt has an important bearing on how we conceptualise existential risks to humanity, including those coming from SAI systems.

Consider, for instance, Nick Bostrom's influential definition of an existential risk as a risk that threatens 'to cause the extinction of Earth-originating intelligent life or to otherwise permanently and drastically destroy its potential for future desirable development' (Bostrom, 2014). Another recent definition of an existential risk is 'an event that brings about the permanent loss of a large fraction of the expected value of humanity' (Greaves, 2024).

In another recent paper focused on AI risks (Cappelen et al., 2025) the authors use the term 'existential risk' as synonymous with 'risk that humanity does not survive'. They clarify that 'not surviving' encompasses 'extinction, near-extinction, and loss of autonomy' (Cappelen et al., 2025).[1] In these definitions and the related debates, it is generally assumed that the focus of discussion is that of an existential risk *to us* and that the extinction of humanity could be just one kind of these risks. However, in these works, it is rarely specified in detail what is meant by 'humanity' or 'us', and consequently whether, say, some forms of AIs could ever be regarded as some of us. Still, if some AIs could be some of us, we may have a different understanding of what we should consider an existential risk for

1 Ord (2020) and Torres (2023) offer other similar definitions of existential risk.

us. For example, cases in which only some SAIs survive might not always count as cases in which *our* extinction occurs.

In this paper, I will focus on the specific use of 'existential risk' in the current debate on existential risks to humanity posed by AI. I will assume that, in this debate, the terms 'humanity,' 'we' and 'us' all refer to the same entities, whatever they may turn out to be. After all, we care about the potential of humanity or humanity's survival at least partly because it's *our* potential and survival. I will assume that this form of self-interest is sufficient to jus-tify our focus on existential risks for humanity. Although the current debate on AI risks is primarily focused on humanity, it is important to notice that there are alternative defini-tions of 'existential risk'. For example, in some contexts, we may want a definition of an exis-tential risk that captures the idea that a risk of this kind is one that severely curtails overall value, not just the value of humanity.[2]

This paper neither addresses the likelihood of existential risk arising from AI systems, nor examines whether SAIs are technically feasible. It does not provide a timeline for their realisation, nor does it make any recommendations for preventing such risks. Instead, my aim is to contribute to the debate by (i) clarifying some of the notions used in it, (ii) connecting it to recent debates on personal ontology, and, lastly, (iii) outlining some interesting consequences for how we think about existential risks posed by AI. More in detail, in Section 1, I outline the main points of an argument for the claim that some forms of AIs constitute an existential risk to humanity. In Section 2, I briefly outline two of the most popular approaches to our nature – animalism and the psychological approach. In Section3, I discuss the theoretical challenges faced by those who suggest that some forms of AI may constitute an existential risk for us. In Section 4, I outline which general conditions an AI should satisfy to be regarded as one of us according to the two approaches introduced in section 2.[3]

2. Existential Risks and Intelligence

There are several arguments for the claim that some forms of AI pose an existential risk to humanity (Cappelen et al., 2025; Müller & Cannon, 2021; Vold & Harris, 2021). One version focused on the possible threats from creating SAI systems goes as follows. There is a non-negligible chance that researchers, or some non-SAI systems, will develop SAIs, AIs whose capacity to achieve their ends or sub-ends far exceeds that of the most intelligent human beings.[4] Unfortunately, being several orders of magnitude better than humans at achieving ends (a form of instrumental rationality), *per se*, does not imply the additional capacity to avoid ends or sub-ends conducive to existential threats to humanity – 'superintelligence' can be prised apart from 'superwisdom' or, at least, from the capacity to avoid outcomes causing the permanent loss of significant expected *human* value.[5] Since an SAI – whether by accident or on purpose – may have ends or sub-ends nefarious to us and given that such an SAI would be several orders of magnitude more intelligent than us, we could hardly stop it, and thus its possible future existence is an existential risk. After all, given the

2 Many thanks to an anonymous referee of this journal for pressing me to clarify the focus of the paper.

3 The issues discussed in this paper should not be confused with similar but relevantly different questions: do AIs of a sophisticated kind have rights, moral status, or (legal) personhood (Basl & Bowen, 2020, Danaher, 2020, Olson, 2022, Schwitzgebel & Garza, 2015, Véliz, 2021)? Can we become AIs or parts of them (E. Olson, 2017; Schneider, 2019)? How can we ensure that AIs' or SAIs' final ends (sometimes called 'values') are aligned with ours (Gabriel, 2020; Ngo et al., 2025)?

4 My understanding of 'intelligence' as used in discussions of AI-related catastrophic or existential risk is drawn from (Bostrom, 2013; Russell, 2014, 1997 and Russell & Norvig, 2020). Some scholars discussing AI systems do not seem to presuppose definitions of 'intelligence' equivalent to the one used in this essay. See (Boden, 2016, Bringsjord & Govindarajulu, 2018, and Hendrycs et al., 2023) for surveys.

5 See (Dung, 2023; Omohundro, 2008; Olson, 2007) for a discussion of the so-called Orthogonality thesis, which is the claim that various levels of intelligence are compatible with more or less any goals. Bostrom clarifies the orthogonality thesis by adding that it does not entail that all of SAIs' means or sub-goals would be entirely imperscrutable. For instance, he suggests that we may predict that some SAIs will have their own survival as at least a means – if, for example, their own persistence would be conducive to their primary or final ends. Or, an SAI may have as a means to its ends the keeping of its ends: 'goal-content integrity' (Bostrom, 2014).

difference in intelligence, we could hardly prevent the realisation of the SAI's existentially threatening ends or sub-ends, should they ever be pursued. Notice that AI systems don't need to be superintelligent to pose an existential threat to humanity. For example, even one or more AI systems that are less than superintelligent but still sufficiently intelligent may constitute an existential risk for similar reasons. I assume that SAI systems constitute the limit case in the intelligence disparity between humans and AIs and thus are theoretically interesting to discuss because of this difference.

The property of being human plays a crucial role not only in the definition of an existential risk for us but also in Bostrom's definition(s) of 'superintelligence'; in fact, both tentative and more refined definitions of this term (except 'collective superintelligence') are spelt out in terms of *human* intelligence. In particular, *superintelligence* is 'any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest' (Bostrom, 2014), *speed superintelligence* is a system that can do all that a human intellect can do but much faster (Bostrom, 2014a), and *quality superintelligence* is a system that is at least as fast as a human mind and vastly qualitatively smarter (Bostrom, 2014a). Bostrom claims that these definitions, all involving a comparison with human capacities, are 'non-committal' concerning how these superintelligences might be implemented (e.g., in an artificial or a biological system) and whether they have phenomenal consciousness or *qualia* (the what-it-is-like to undergo a specific experience from a first-person perspective). In this paper, I will assume that intelligence is one of *our* properties, that it is possessed to various degrees, and that it is, or essentially involves, at least the capacity (or group of capacities) to consistently achieve ends through a selection of proper means, given an agent's contingent conditions. An SAI discussed in this paper will be understood as an artificial entity endowed with a level of instrumental intelligence far superior to that of the most gifted past or present human being. Now, what are some of the most popular theories in the current ontological debate on our nature and how can they affect our thinking of what constitutes an existential risk for us?

3. Our Ontological Category

An ontological theory of our nature is understood here as a systematic attempt to answer the following questions: What is the ontological nature of entities like *you* and *me*? Under what ontological category or categories (concrete particulars, psychological entities, narra-tives, etc.) should individuals like us be classified?[6] One deceptively straightforward reply to these questions is that we are human beings. This idea can be further specified by saying that we are human biological organisms, specifically organisms of the species *Homo sapiens* – a view called *organism-animalism*, or simply *animalism*.[7] On some versions of animalism, we are each *essentially* a living human organism – that is, we each instantiate the property of being a living human organism whenever we happen to exist. Another version of ani-malism claims that each one of us is essentially a human organism *simpliciter* – thus omit-ting 'living'. On this second view, each of us would continue to exist even when the organ-ism each one of us is identical with has ceased functioning (e.g., after its metabolism has ceased supporting those life processes that ensure the persistence of its biological functions over time). The main difference between these two versions of animalism concerns

6 I follow Olson (2007), Sauchelli (2018) and many others in regarding ontological theories of our nature as conceptually distinct from theories of *personal* identity, although the two kinds of theories are related. I do not consider in this paper a further complication deriving from a possible ontological distinction between, on the one hand, *you* (singular) and *me* and, on the other, *you* (plural) and *us*. In fact, I will regard 'we' and 'us' as the mereological sum of me and each of you, not an ulterior, complexly structured social object. See (Epstein, 2018) for a clear introduction to the current debate in social ontology. Other approaches are possible (e.g., Dung, 2024): 'since humanity is an abstract entity, […]'), but do not substantially change the main points of this paper.

7 There are complex issues concerning the ontology of (i) organisms, (ii) biological organisms, and (iii) biological individuals, not to mention those concerning the metaphysics of species. See (Clarke, 2010; Godfrey-Smith, 2015; Guay & Pradeu, 2016; Olson, 2020) for discussion. In this paper, I will assume that biological organisms, like human beings, form a proper ontological category. Olson (2007) and (Bailey & Elswyk, 2021) discuss several arguments for animalism.

whether a human organism must be 'living' to count as one of us.[8] According to the latter view, a corpse may still count as one of us. In contrast, on the former, when a human organism ceases to function, it simply ceases to exist – that is, it does not enter a new phase of existence, the 'dead phase', but merely is annihilated.

Animalism can be further specified by including an account of organisms' synchronic and diachronic identity conditions. Although this is a controversial issue in the philosophy of biology and biological theory, I will assume that, at least for human beings, there are sufficiently clear cases (e.g., adult human organisms) that can be reasonably taken as the central exemplars of their kind. Such exemplars may provide the basis for formulating sufficiently coherent and informative identity conditions.[9] For instance, (Guay & Pradeu, 2016), elaborating on (Hull, 1978) and (Hull, 1992), make the point that the persistence of a human organism is not metaphysically dependent on the continuous persistence of an immutable core or an unchanging set of parts but rather depends on the degree of continuity that the organism's internal organisation has. To the extent that the internal organisation of the organism persists to a sufficient degree, the organism persists as well (Guay & Pradeu, 2016).

Several researchers from different disciplines seem to adopt versions of animalism. For instance, according to Edward O. Wilson, '[w]e are a biological species arising from Earth's biosphere as one adapted species among many [...]' (Wilson, 1978). In his recent *Homo Sapiens Rediscovered*, Paul Pettitt writes: '[w]ho are we? How do scientists define *Homo sapiens*, and how does our species differ from the extinct humans that came before us [...]' (Pettitt, 2022). Perhaps this acceptance is because many think that this is the sole scientifically acceptable thesis about our nature – the only scientifically viable alternative to views suggesting that we are souls, compounds of souls and bodies, or supernatural entities of some other kind. For instance, Geerat Vermeij states, '[i]f there is one thing science has taught us about ourselves, it is that we humans are animals' (Vermeij, 2023). However, this reason for adopting animalism is not compelling. In fact, the so-called psychological approach, combined with a scientifically acceptable view of the nature of the mind, is a plausible alternative. According to theories belonging to this approach, we are psychological entities (e.g., minds or selves), and the existence of beings of this kind is conceptually compatible with the presuppositions and findings of various scientific disciplines.

The psychological approach differs from animalism primarily because the identity conditions of a human organism are not the same as those of psychological entities like 'minds' or 'selves'.[10] For example, an organism kept alive by machines but no longer capable of remembering, having conscious experience, and other mental states would seem to be a case in which a psychological entity, but not the related organism, has ceased to exist. In addition, what seems to be necessary to maintain a mind or self into existence (e.g., a functioning nervous system) need not coincide with what is required for the continuous existence of the whole organism. If there are cases in which a human organism persists but its related mind does not, and *vice versa*, then minds and human organisms are distinct entities. Still, the existence and persistence of a mind may depend on specific physical properties – a view amenable to several scientific theories.[11]

Intuitions related to brain-transplant scenarios in the literature on personal identity seem to support versions of the psychological approach (Parfit, 1984; Sauchelli, 2018). In particular,many people seem to believe that if a brain transplant from one body to another were to ensure the persistence of our mind in the receiving body, a brain trans-plant may result, at least theoretically, in an exchange of body. This intuition suggests that we associate our identity (not just our personal identity) with our mind rather than with

8 See (Sauchelli, 2017) for discussion.

9 See note 7 for references.

10 For instance, see (Shoemaker, 2011) for a contemporary Neo-Lockean view. According to Shoemaker, we are entities having mental states whose causal profiles determine our identity conditions. Although we are constituted by animal organisms, we are not identical to them.

11 See (Birch, 2024; Feinberg & Mallatt, 2018) and (Levin, 2022) for relevant discussion.

the whole organism in which our brains are located. The supporter of the psychological approach to our nature would claim that intuitions about these cases show that (i) a biological organism's and a psychological entity's conditions of identity do not coincide, and (ii) we (metaphysically or numerically) identify ourselves with a psychological entity.

The psychological approach to our nature can be further specified along different dimensions of variation. More specifically, psychological theories may differ with respect to:

1. the kind of mental states that are regarded as constitutive of our identity and their continuity over time (e.g., episodic memory, consciousness, moral character, intentions and plans, and so on);
2. whether our mental states are supposed to satisfy a quantitative condition (e.g., that at least 50% of the psychological connections that normally hold between two subsequent days should hold) to sustain our identity;
3. whether our mental states are supposed to have a qualitative structure — for instance, whether they should form clusters of relations that, in turn, form (1) a narrative, (2) a volitional, or (3) a (moral) agential structure;
4. the way in which such mental states are connected (e.g., by causal chains or by being instantiated in the same brain).[12]

For the purposes of this paper, I will consider a generic version of the psychological approach, one compatible with any (or almost any) of the above further specifications.

4. Existential Risks and Us

According to the argument presented in section 1, the existential risks associated with the development of SAI systems are partly dependent on their highly instrumental effectiveness and on the possibility that their goals or sub-goals may not be aligned with humanity's. Their superiority in effectiveness compared to us significantly contributes to justifying a certain level of concern when assessing their potential to pose existential risks to humanity. Still, participants in the debate on AI-existential risks do not always offer many details on their understanding of 'humanity' or 'us.'

In the first pages of Bostrom's *Superintelligence*, the referent of 'we' seems to be humans, intended as *Homo sapiens* in a biological or evolutionary sense (Bostrom, 2014b).Bostrom is not the only prominent writer to make this implicit assumption. For instance, Stuart Russell and Peter Norvig write: 'We call ourselves *Homo sapiens* – man the wise – because our intelligence is so important to us' (Russell & Norvig, 2020).In a recent presentation of the 'control problem,' Vincent Müller writes that it concerns 'how we humans can remain in control of an AI system once it is superintelligent' (Müller & Cannon, 2021) Still, as the introduction noted, an existential risk to humanity seems to have, in the writings of some scholars, a range of application that includes risks to other '[…] Earth-originating intelligent life' of a kind similar to *Homo sapiens*. For instance, Toby Ord claims that: [i]f we somehow give rise to new kinds of moral agents in the future, the term 'humanity' in my definition should be taken to include them (Ord, 2020). Similarly, Hilary Greaves suggests that: 'If *Homo sapiens* underwent continued evolution to such an extent that our successors came to count as members of a distinct biological species, that would not in and of itself be cause for concern' (Greaves, 2024).So, although participants in the debate on AI existen-tial risks generally focus on risks to humanity, it is not always clear whether, in the writing of some of them, the referents of 'humanity' or 'us' include only *Homo sapiens* or not. If par-ticipants in the debate on existential risks regard as plausible to include as referents of 'us' or 'we' human beings (*qua* members of a species) *and* other entities(because of some rela-

12 See Sauchelli (2018) for details. For reasons of space, I have excluded some theories of our nature (e.g., nihilistic views).

tional properties with humans), then we would need a more explicit account of this extension.

In general, one of the theoretical challenges for these scholars is that of providing a conception of 'us', 'we', etc., that is inclusive enough to encompass some of our descendants but that is exclusive enough to consider some (dangerous) forms of AI as not 'us'. Such an expanded definition may be preferable because (i) in a sufficiently long timespan, individuals evolutionarily related to our species may or will likely evolve new traits such that identifying them as *Homo sapiens* would be a classificatory mistake, and (ii) not all cases in which, following this evolution, *Homo sapiens* go extinct would plausibly amount to our extinction. For instance, we may argue that if a newly evolved species is properly connected (e.g., through sufficiently and adequately causal and spatiotemporal connections of the kind that biologists would rationally converge in regarding as an ancestry relation) with *Homo sapiens* and exemplars of this species maintain most or all of the traits that we regard as crucial for sustaining our identity, then we may regard this evolution as a way for *us* to continue to exist.

To illustrate this point, suppose that, due to future environmental damage, all our successors will develop inheritable traits that future biologists would rightly regard as indicative of a new species. Suppose that members of this new species will be able to understand and appreciate most of our cultural productions, speak some of our languages, develop new technologies improving on ours, and so on. In short, suppose that these descendants will still have minds that are relevantly similar to ours. Would that outcome be a way for us to go extinct? I suspect that at least some participants in the debate would say that this would *not* be a way for us to go extinct. I am not here suggesting that all participants in the debate would be willing, upon reflection, to extend their conceptions of 'humanity' or 'us' to also include members of other species. Still, extending the notion of 'humanity' to include at least some of our evolutionary descendants is a plausible theoretical move that some theorists debating AI-existential risks seem to be already making. To evaluate such a move, let us call the disjunctive property of being *Homo sapiens* or any other kind of biological organism properly related to them, 'H'. This disjunctive property need not be understood as involving an exclusively biological relation but may include other elements of the kind mentioned before — for example, being a biological organism that is also our 'cultural successor' in virtue of retaining those properties required to appreciate and develop our cultural and technological productions. If this is plausible, we can now define the possible victims of AI-existential risks for us as those entities that have or will have H. Is this conception compatible with the views on personal ontology discussed above?

First, most psychological theories would agree that we presently instantiate H. More specifically, they may suggest that H is one of our contingent properties since H is not a property on the loss of which one of us would necessarily cease to exist. One problem with combining this approach with a definition of an existential risk that defines 'humanity' or 'us' in terms of H is that (i) since H would be at most one of our contingent properties, an event in which we cease to possess H would not necessarily be an event in which we cease to exist or an event that may curtail our potential, so (ii) definitions of existential risk which used H to determine their scope would be at least inaccurate, or even wrong. As an illustration of this point, consider those transhumanists – presumably including Bostrom – who do not believe that being a biological organism (or being embodied in one) is a necessary condition for one of us to exist or persist.[13] These transhumanists should not adopt the above definition of an existential risks based on H. In fact, if they believe that we may persist as uploaded individuals in some super-computer, or by having our bodies slowly replaced (fully or almost fully) by non-biological matter, they should also believe that we would persist even without subsequently instantiating H.[14] If so, having H should not be what exclusively determines the extension of the subjects of existential risk in the debate, if

13 See Olson (2022) for the metaphysical presuppositions of some transhumanist beliefs about persistence.

14 See Chalmers (2010) for a discussion of how the metaphysics of mind may relate to mental uploading.

only because a scenario in which no entities instantiate H is not necessarily a scenario in which we cease to exist. In conclusion, some transhumanists and those supporters of the psychological approach who do not believe that having a (fully) biological organism is necessary for our persistence would have to choose between (i) the belief that they could survive without having H and (ii) a definition of existential risk based on the possession of H. Of course, (i) is a belief that partly depends on the preferred metaphysics of mental states.

Other supporters of the psychological approach may claim that having the property H, or a slightly modified version of it, could be a condition relevant to our persistence and thus to the extension of 'us'. For instance, some versions of this approach may embrace a metaphysics of mental states according to which the biological embodiment of a kind falling under the scope of H is a nomologically essential condition for having those mental states that sustain our identity over time. In other words, on some versions of the psychological approach, having those mental states that determine our identity requires specific biological properties. A further refinement of this kind of psychological theory may thus encompass a slight modification of H to include entities with those biological features that, according to a suitable theory of the mind, are at least nomologically required for having those mental states that are, in turn, required for us, *qua* minds or selves, to exist and persist. For example, supporters of this version of the psychological approach may amend H to include as an additional or alternative disjunct the property of having a biological nervous system that supports specific required mental structures (H*).

In reply, it may be argued that possession of H* is not what properly individuates the relevant class of 'us': having H* may be only contingently required for having the relevant mental states. That is, we may argue that H* is a property we must possess solely because mental states must be realised in biological organisms of a particular kind (given our cur-rent laws of nature).[15] However, if they could be otherwise realised, then possession of H* would not serve to individuate the proper extension of 'us' as the subjects of an existential risk. Still, supporters of the relevant versions of the psychological approach may just be content with extensional accuracy in this case – provided that H* is properly defined and that the extension of 'us' relevant for defining existential risk would coincide with those entities supporters of the approach identify as 'us'.

In general, adopting a psychological approach requires modifying the understanding of the extension of the subjects of an existential risk in a way that can fit the specific theory's position along the approach's dimensions of variation. For example, suppose a theory of our nature claims that we are minds whose persistence is determined by a certain level of psychological continuity over time, and such continuity depends on a certain number of psychological connections over time. In that case, a definition of an existential risk to humanity should imply that the subjects of risk are all those entities satisfying these conditions. An interesting theoretical consequence of this approach is that in addition to the usual list of existential risks to humanity (Ord, 2020), we may also have to include events causing, for example, a mass interruption of the causal chains ensuring our memories – insofar as memory-continuity is regarded as essential to our survival. A better understanding of the ontological category to which we belong may thus reveal other ways in which our existence may be at risk.

What about the other prominent theory of our nature – animalism? Animalists who adopt the definition of an existential risk proposed by Bostrom would have to amend their view of our nature in a way not entailing that if we cease to be *Homo sapiens*, we will also thereby cease to exist. On a version of animalism claiming that we are living biological organisms, this modification would be minimal: the existential risk theorist of an animalist persuasion may rephrase the proposed conditions of identity without referring to our current species. For example, the amended form of animalism may simply include H in its account of our nature and claim that to be one of us is to be a living organism that is a member of *any or some* species properly related to *Homo sapiens*. This relation between species

15 See Kim (1992) for a classic paper on realisation.

can be specified to include the condition that, to be properly related to us, a subsequent species of *Homo* should be evolutionarily connected to our species and retain certain key aspects of our genus (e.g., bipedal locomotion, tool use and modification of the environment, complex social organisation, cultural transmission mechanisms, capacity to use languages, and so on). On this amended form of animalism, future biological organisms of different species but still belonging to the genus *Homo* may be 'properly related' to us and thus count as some of us even if they are not *Homo sapiens*. However, it remains an open question whether this form of animalism ('H-animalism') would be plausible, primarily because this revised form of 'H'— a disjunctive, potentially open property — may be perceived as unsuitable for determining our nature.

A different kind of problem for the compatibility of animalism with the apparent theoretical commitments of at least some of those debating AI-existential risks concerns the likely divergence of intuitions about our persistence and the theoretical outcomes of H-animalism. Again, suppose that the only way for us to persist may involve a radical alteration of (a significant part of) our internal biological processes, such that it would leave only the upper parts of our brains or our nervous systems intact. For instance, suppose that gradually replacing the rest of our bodies with non-biological material will be necessary to sustain a certain acceptable level of brain function due to the sudden deterioration of the environment. These 'Earth-originating' entities may even subsequently decide to revert to a biological embodiment similar to our current one. Now, an H-animalist would have to claim that this process would constitute an existential risk – more specifically, it would involve our extinction, even in case the process were successful in transforming our bodies. In fact, the discontinuity in the biological organisation of our organisms would be so significant that it is not plausible to claim that the same organisms would persist after the procedure. However, I suspect many of those concerned with existential risks would not regard this scenario as equivalent to our extinction. Those sharing this latter intuition would better adopt some version of the psychological approach as a theory of our nature. In conclusion, some versions of the psychological approach seem to be more amenable to the intuitions of several scholars working on existential risks to humanity. If that is correct, at least these scholars should refrain from suggesting that some forms of AIs threaten us as human beings – intended as entities essentially belonging to the species *Homo sapiens*.

The following section expands on some of the previous considerations. It focuses on another clarification that debates on personal ontology may bring to the debate on existential risks coming from AI: under what conditions would an AI or SAI be one of us (if at all)?

5. Can an AI Be One of Us?

Intelligence is what *we* seem to have in common with an SAI. The main difference is that an SAI will have a significantly more effective form of instrumental rationality. Having superintelligence and possessing instrumental rationality at our current level do not seem to be different in kind, although it is unclear what would be required to have a superintelligent level of instrumental rationality – for example, whether it would require the development of new types of mental abilities presently unknown. Leaving this last point aside, having such a 'super' level of intelligence is the only necessary condition for being an SAI system which is currently stipulated explicitly – in addition to being 'artificial'. In this section, I will explore what some versions of animalism and the psychological approach discussed before implying for identifying an SAI with one of us. As mentioned, the discussion here is focused on SAI systems partly because they seem to represent a limit case in terms of a relevant difference (i.e., level of intelligence) from us. If an SAI could be one of us, then presumably the case for less than superintelligent AIs would be even less controversial.

According to animalism and H-animalism, a non-biological SAI system cannot be one of us. As a consequence, these theories of our nature imply that takeovers by non-biological SAIs and similar scenarios involving the extinction or curtailed future potential of *Homo sapiens* or H-holders should be regarded as existential risks for *us*. In other words, if the arti-

fificial agents are not also *Homo sapiens* or H-holders, the forms of animalism discussed here imply that all AI-existential risks threatening the long-term potential of entities essentially having H would be existential risks for us.

As we have seen, versions of the psychological approach vary depending also on which metaphysics of mental states it is combined with. For example, consider a version of this approach according to which we are psychological entities whose identity conditions over time require a certain degree of psychological continuity and where this essentially in-volves continuity of memories, intentions, plans, character traits, and the presence of (phenomenal) consciousness. If such a view is paired with a metaphysics of mental states according to which phenomenal consciousness or other relevant mental states nomologically require a specific biological substratum (e.g., the so-called neural correlate of consciousness), albeit not necessarily a human one, then SAI systems may count as some of us only if they have consciousness and its required biological substratum. For example, creatures with superintelligent levels of instrumental rationality and phenomenal consciousness, supported by the proper biological material, might count as some of us. As a consequence, a scenario in which such creatures eliminate or dominate all *Homo sapiens* need not count as an existential risk to (all of) us. It may indeed be a catastrophe if their actions would bring about the violent destruction of all *Homo sapiens* (intended as biological organisms), but the takeover need not be hostile. However, this sort of hybrid AI system having a biological substratum is generally not what is considered in the recent debate on AI-existential risks. In fact, scholars usually assume that AI systems that pose threats to humanity will likely not be realised in biological systems of any kind. If that were the case, then versions of the psychological approach suggesting that we are minds essentially realised in a biological substratum (say, an organic brain) would not regard AI systems that do not have such a substratum as some of us.

If other theories of mental states and consciousness are true – say, some version of functionalism – it may not be necessary for us to persist that we possess a biological substrate. To the extent that what is regarded as crucial for our existence and persistence (say, consciousness, the continuity of some mental states and that of interconnections among them) is present also in AIs or SAIs, they may count as some of us – perhaps as our exceptionally gifted brethren. In scenarios in which SAIs with these features uniquely persist on Earth, their coming into existence itself may or may not count as a catastrophic risk, primarily depending on the reasons for this unicity. Perhaps these new entities tried to help *Homo sapiens* in the aftermath of an inevitable environmental catastrophe but failed to save them. Perhaps an engineered pandemic killed off all *Homo sapiens* through no fault of the new SAIs. We may even imagine scenarios in which our long-term potential or the expected value of humanity may require some sort of substitution of *Homo sapiens* with new SAIs. In these cases, supporters of this functionalist version of the psychological approach may have reasons to regard these scenarios as at least better for them and humanity than those involving solely the extinction of all *Homo sapiens*.

Conclusions

While this paper did not argue for a theory of our nature rather than another and thus of what should count as an existential risk for us, its contribution lies primarily in the clarification of key concepts within the debate on AI-existential risks for humanity. In addition, my purpose was not to discuss a definition of existential risk that would include a definition of humanity capturing (directly) what we value or what is of value (Finneron-Burns, 2024; Railton, 2023). Rather than defining 'humanity' on the basis of, say, what we value about it or its value, I started from theories of our ontology. Considerations of value may be relevant to some of these theories, for example, to some psychological theories in defining what determines our identity conditions over time and, thus, indirectly, what 'us' stands for. However, considerations of value were not the primary or direct criteria for selecting a definition of an existential risk for us.

A recent attempt that centrally addresses the definition of humanity and what we value

about it is (Finneron-Burns, 2024). In particular, Elizabeth Finneron-Burns has recently argued that, in the context of discussing human extinction, we should adopt a definition of humanity as 'human bodies with reason, culture, and flourishing' as it better reflects what we value about humanity. An existential risk for us that follows this definition would be a risk that necessarily threatens, among other things, the existence of *Homo sapiens*. Although this is a different project from mine, Finneron-Burns' defence of this definition does not strike me as persuasive – even within the context of providing a definition of humanity that captures what we value. In particular, what does not strike me as persuasive is the defence of the claim that we value if 'human bodies' themselves are those achieving things and that this should inform the definition of humanity (Finneron-Burns, 2024). More specifically, she suggests that we should retain the point that humanity must include hu-man bodies because it may matter *subjectively* – it may matter to us that it is human bodies that continue to exist or achieve things. This explanation also does not seem satisfactory because, presumably, there should be a sufficiently strong motivation that such kinship should matter to us for it to be a properly motivating reason for rational practical choices (e.g., choosing certain policies over others).

In conclusion, this paper has explored the interplay between ontological theories of our nature and the current debate on existential risks posed by AI. Ultimately, this exploration has revealed that different ontological theories of our nature determine which scenarios are identified as existential threats to us and provided some guidance on understanding under what conditions AI systems can be regarded as some of us. Although I did not directly argue in favour of any position in the debate, I will conclude by mentioning that I am doubtful of the claim that some AI systems (at least in their current variety) can be regarded as some of us primarily because I am not yet convinced that instances of phenomenal consciousness of the kind we deem as characterising our identity are nomologically possible in digital or non-biological forms. This conclusion would not entail that, should AI systems develop some sort of non-human form of consciousness or satisfy other requirements for having moral status, we should not act morally towards them. Arguably, at least some kinds of animals deserve moral recognition although they are not part of humanity – similarly, AI systems may deserve an equal amount of moral recognition even if they are not some of us.

References

Bailey, A., & Elswyk, P. (2021). Generic Animalism. *Journal of Philosophy*, *118, 8*, 405–429.

Bales, A., D'Alessandro, W., & Kirk-Giannini, C. D. (2024). Artificial Intelligence: Arguments for Catastrophic Risk. *Philosophy Compass*, *19*, 2.

Basl, J., & Bowen, J. (2020). AI as a Moral Right-Holder. In M. et alia Dubber (Ed.), *Oxford Handbook of Ethics of AI* (pp. 289–306). Oxford University Press.

Birch, J. (2024). *The Edge of Sentience*. Oxford University Press.

Boden, M. (2016). *AI. Its Nature and Future*. Oxford University Press.

Bostrom, N. (2012). The Superintelligent Will. *Minds and Machines*, *22*, 71–85.

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. *Global Policy*, *4, 1*, 15–31.

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Bringsjord, S., & Govindarajulu, N. S. (2018). Artificial Intelligence. *Stanford Encyclopedia of Philosophy*.

Cappelen, H., Goldstein, S., & Hawthorne, J. (2025). *AI Survival Stories. Philosophy of AI*. Vol. 1, 59–70.

Chalmers, D. (2010). The Singularity. *Journal of Consciousness Studies*, *17*, 7–65.

Clarke, E. (2010). The Problem of Biological Individuality. *Biological Theory*, *5*, *4*, 312–325.

Danaher, J. (2020). Welcoming Robots into the Moral Circle. *Science and Engineering Ethics*, *26, 4*, 299–309.

Dung, L. (2023). Current Cases of AI Misalignment and Their Implications for Future Risks. *Synthese*,

202, *138*.
Dung, L. (2024). *The Argument for Near-Term Human Disempowerment through AI*. *AI & Society*.
Epstein, B. (2018). *Social Ontology*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/social-ontology/
Feinberg, T., & Mallatt, J. (2018). *Consciousness Demystified*. The MIT Press.
Finneron-Burns, E. (2024). Humanity': Constitution, Value, and Extinction. *The Monist*, *107*, 99–108.
Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*, 411–437.
Godfrey-Smith, P. (2015). Individuality and Life Cycles. In A. Guay & T. Pradeu (Eds.), *Individuals Across the Sciences* (pp. 85–102). Oxford University Press.
Greaves, H. (2024). Concepts of Existential Catastrophe. *The Monist*, *107*, 109–129.
Guay, A., & Pradeu, T. (2016). To Be Continued. The Genidentity of Physical and Biological Processes. In A. Guay & T. Pradeu (Eds.), *Individuals Across the Sciences* (pp. 317–347). Oxford University Press.
Hendrycs, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks*, *v6*. Arxiv. https://arxiv.org/abs/2306.12001
Hull, D. (1978). A Matter of Individuality. *Philosophy of Science*, *45*, *3*, 335–360.
Hull, D. (1992). Individual. In E. F. Keller & E. Lloyd (Eds.), *Keywords in Evolutionary Biology* (pp. 181–187). Harvard University Press.
Kim, J. (1992). Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research*, *52*, 1–26.
Levin, J. (2022). *The Metaphysics of Mind*. Cambridge University Press.
Müller, V. (2020). Ethics of Artificial Intelligence and Robotics. *Stanford Encyclopedia of Philosophy*.
Müller, V., & Cannon, M. (2021). Existential Risk from AI and Orthogonality. *Ratio*, *35*, *1*, 25–36.
Ngo, R., Chan, L., & Mindermann, S. (2025). The Alignment Problem from a Deep Learning Perspective. *arXiv:2209.00626v8*. https://doi.org/https://doi.org/10.48550/arXiv.2209.00626
Olson, E. (2007). *What Are We?* Oxford University Press.
Olson, E. (2017). The Central Dogma of Transhumanism. In B. Berčič (Ed.), *Perspectives on the Self* (pp. 35–58). University of Rijeka.
Olson, E. (2020). What Is the Problem of Biological Individuality? In A. S. Meincke & J. Dupré (Eds.), *Biological Identity: Perspectives From Metaphysics and the Philosophy of Biology* (pp. 63–85). Routledge.
Olson, E. T. (2022). The Metaphysics of Transhumanism. In K. Hu?bner (Ed.), *Human: A History (Oxford Philosophical Concepts)* (pp. 381–403). Oxford University Press. https://philarchive.org/rec/OLSTMO-12
Omohundro, S. (2008). Basic AI Drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial General Intelligence: Proceedings of the First AGI Conference* (Vol. 171, pp. 483–492). IOS Press.
Ord, T. (2020). *The Precipice*. Hachette Books.
Parfit, D. (1984). *Reasons and Persons*. Clarendon Press.
Pettitt, P. (2022). *Homo Sapiens Rediscovered*. Thames & Hudson.
Railton, P. (2023). The Normative Significance of Humanity. In S. Buss & L. N. Theunissen (Eds.), *Rethinking the Value of Humanity* (pp. 273–290). Oxford University Press.
Russell, S. (1997). Rationality and Intelligence. *Artificial Intelligence*, *94*, 57–77.
Russell, S. (2014). Rationality and Intelligence: A Brief Update. In V. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 7–28). Springer.
Russell, S. (2019). *Human Compatible*. Penguin.
Russell, S., & Norvig, P. (2020). *Artificial Intelligence* (4th ed.). Pearson.
Sauchelli, A. (2017). The Animal, the Corpse, and the Remnant-Person. *Philosophical Studies*, *174*, 205–218.
Sauchelli, A. (2018). *Personal Identity and Applied Ethics*. Routledge.
Schneider, S. (2019). *Artificial You*. Princeton University Press.
Schwitzgebel, E., & Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Studies in Philosophy*, *XXXIX*, 98–119.
Shoemaker, S. (2011). On What We Are. In S. Gallagher (Ed.), *Oxford Handbook of the Self* (pp. 352–371). Oxford University Press.
Torres, E. (2023). Existential Risks. *Inquiry*, *66*, *4*, 614–639.

Turchin, A., & Denkenberger, D. (2020). Classification of Global Catastrophic Risks Connected with Artificial Intelligence. *AI & Society*, *35*, 147–163.

Véliz, C. (2021). Moral Zombies. *AI & Society*, *36*, 487–497.

Vermeij, G. (2023). The Origin and Evolution of Human Uniqueness. In H. Desmond & G. Ramsey (Eds.), *Human Success* (pp. 91–116). Oxford University Press.

Vold, K., & Harris, D. (2021). How does Artificial Intelligence Pose an Existential Risk? In C. Véliz (Ed.), *The Oxford Handbook of Digital Ethics*. Oxford University Press.

Wilson, E. O. (1978). *On Human Nature*. Harvard University Press.

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom & M. Ćirković (Eds.), *Global Catastrophic Risk* (pp. 308–345). Oxford University Press.