Philosophy & Technology (2024) 37:16
https://doi.org/10.1007/s13347-024-00706-2

RESEARCH ARTICLE

Human Extinction and AI: What We Can Learn from the Ultimate Threat

Andrea Lavazza[1,2] · **Murilo Vilaça**[3]

Received: 24 July 2023 / Accepted: 12 January 2024 / Published online: 1 February 2024

Abstract

Human extinction is something generally deemed as undesirable, although some scholars view it as a potential solution to the problems of the Earth since it would reduce the moral evil and the suffering that are brought about by humans. We contend that humans collectively have absolute intrinsic value as sentient, conscious and rational entities, and we should preserve them from extinction. However, severe threats, such as climate change and incurable viruses, might push humanity to the brink of extinction. Should that occur, it might be useful to envision a successor to humans able to preserve and hand down its value. One option would be to resort to humanoid robots that reproduce our salient characteristics by imitation, thanks to AI powered by machine learning. However, the question would arise of how to select the characteristics needed for our successors to thrive. This could prove to be particularly challenging. A way out might come from an algorithm entrusted with this choice. In fact, an algorithmic selection both at the social and at the individual level could be a preferred choice than other traditional ways of making decisions. In this sense, reflecting on human extinction helps us to identify solutions that are also suitable for the problems we face today.

Keywords Existential risk · Human value · Humanoid robot · Algorithmic enhancement · Future of humanity

✉ Andrea Lavazza
lavazza67@gmail.com

[1] University of Pavia, Department of Brain and Behavioral Science, Pavia, Italy

[2] Centro Universitario Internazionale, Arezzo, Italy

[3] Department of Human Rights, Oswaldo Cruz Foundation (Fiocruz), National School of Public Health (ENSP), Rio de Janeiro, Brazil

1 Introduction: Current Existential Risks

That the human species might become extinct is a prospect that has only recently started to surface often, although the idea of a cyclical destruction and rebirth of a civilization – if not the universe as a whole – is an ancient one, recurring in many different cultures. For example, dates marking a new millennium, such as the year 1000, were often viewed as potentially signifying the end of the world. That being noted, it seems to be safe to say that the idea of human extinction spread mainly in the twentieth century because of the fear of a thermonuclear war, which could have caused the destruction of life on Earth or at least the death of most if not all human beings.[1] More recently, climate change – caused by human activity – has shown that there is a risk of the Earth becoming uninhabitable for humans. The pandemic outbreak due to COVID-19 has also revived fears[2] that a deadly virus will spread to the point of annihilating the human species.[3] In Ord's terms, it seems that we are on the edge of a precipice and that the future of humanity is extremely uncertain (Ord, 2020).

However, it should not be forgotten that despite the awareness of the risks humanity is running today, some positions suggest that we can look to the future of humankind with optimism.[4] According to Bostrom, for example, it is dangerous to be alive, but luckily not all risks are serious (Bostrom, 2002). In his view, the magnitude of risks can vary in scope (e.g., "the size of the group of people that are at risk"), intensity (e.g., "how badly each individual in the group would be affected") and probability (e.g., "the best current subjective estimate of the probability of the adverse outcome") (Bostrom, 2002, 1).

Now, there are risks whose scope and intensity can reach catastrophic, unbearable, and irreversible dimensions. The concept that has been used by some authors to define such situations is that of *existential risks* (ERs). Proposing a distinction among six types of risk based on their scope and intensity, Bostrom defines existential risk as a terminal global risk. Unlike personal, local and global endurable risks, ERs are the ones at the highest level in terms of both scope and intensity. That is, an ER is one whose adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential (Bostrom, 2002). An ER is one where humankind as a whole is imperilled.

[1] The so-called Doomsday Clock, created in 1947 by the members of the Bulletin of the Atomic Scientists, signals the likelihood of a man-made global catastrophe. At first, the only threat considered was that of a nuclear war, but recently climate change has been added. As of January 2023, the clock marks 90 s before midnight, i.e., the hypothetical end of the world.

[2] Mary Shelley's famous work *The Last Man* (1826) describes the extinction of humankind through epidemics of communicable diseases, especially the plague.

[3] Although vaccines developed in an unexpectedly short time appear to have greatly diminished the threat of the SARS-CoV-2 virus, the presence of other similar and potentially equally lethal viruses has been documented.

[4] Take, for example, a widely acclaimed book such as *Homo Deus* by Yuval Noah Harari; cf. also Vilaça and Lavazza (2022).

So, for Ćirković, Sandberg and Bostrom, "ERs include global nuclear war, the collision of the Earth with a 10-km sized (or larger) asteroidal or cometary body, intentional or accidental misuse of bio- or nanotechnologies, or runaway global warming" (Ćirković et al., 2010, 1495).

As can be seen, ERs are mostly recent: to a certain extent, they were metaphorically triggered with the first nuclear bomb (Bostrom, 2002). According to Moynihan, existential risks have become the target of an emerging field of scientifically serious studies, but the dynamic of incremental future guidance is not exactly new, dating back to the Enlightenment (Moynihan, 2020). From this perspective, the question of risk is, in a sense, a matter that involves rational faculties, such as prediction, intervention, mitigation, and self-responsibility for the future of humanity.

Persson and Savulescu have long been pointing out a paradox of our time: the technoscientific advance that radically improved human life on Earth is the same that might contribute to its total extinction (Persson & Savulescu, 2008, 2012). Their central theses are that our moral psychology does not evolve fast enough; that we are unfit for the future; that social measures are insufficient and that a moral bioenhancement is necessary and urgent so that we can avoid the risk of extinction caused by human action itself (Persson & Savulescu, 2012).

In the face of these threats, which are largely the result of human activity, the prospect of *Homo sapiens*' extinction suggests the need to think about how we might act pre-emptively. In this article, we will focus on a necessarily hypothetical scenario that is made as realistic as possible. However, our purpose is primarily to present some ethical considerations rather than to provide detailed technological detail related to digital duplicates and machine learning application.

The premise, therefore, is that there is an impending threat and that there is a will to address the issue of our potential extinction (cf. MacAskill, 2022). Faced with a worsening of climatic conditions, or the repeated occurrence of epidemics caused by increasingly aggressive and incurable viruses, as well the possible misuses of science and technology (such as the creation of lethal biological agents in the lab; the uncontrolled release of genetically modified species into the environment or unsuccessful attempts to reverse climate change—for example, the excessive release of silver iodide into the atmosphere to seed clouds), it might be worth starting to contemplate a way to pass what is the best in the human species to what might be our "heirs", i.e., entities which we will describe below.

According to pessimists about the development of artificial intelligence (cf. Bostrom, 2014; Russell, 2019; Tegmark, 2017), another major threat to humanity may come from out-of-control machine superintelligence. If such AI systems were capable of destroying humanity, then surely AI could also override any human plans for the creation of AI successors. In our perspective of a possible solution to human extinction, the answer to this objection can be twofold.

On the one hand, the possibility of a super-intelligent AI proving destructive and capable of taking over the whole of humanity depends on the stage at which it will implement its plan with respect to the process of creating our artificial successors. If artificial successors were ready in adequate numbers, their network might be able to thwart the attempt of a 'rebel' AI. But of course, with these scenarios, we are

moving on even more speculative ground than that of individual artificial successors, and so it is difficult to make plausible predictions.[5]

On the other hand, it could be argued that the super-intelligent digital AI that would take the place of all carbon individuals would be a 'super-successor' to humanity, as it was originally created and educated by humans. We cannot know whether this 'super-successor' would be the ideal successor we would like to have, but it would certainly incorporate much of what humans have been and done. We can think of a story similar to that of Cain who kills his brother Abel and then gives rise to an offspring in which there are good and bad individuals regardless of their predecessor, although there is a difference between biological and digital entities.

So, the underlying scenario to our project can only be that of a slow-onset and slow-development threat, since sudden events such as a massive thermonuclear war would not leave time to take effective measures to deal with human extinction, unless we prepared in advance for a sudden catastrophic event. In that case the quantity of successors that we could make active would inevitably be small. The number of successors that we would like to produce and somehow keep ready for a sudden catastrophic event might reflect, if the choice is as rational as possible, the value we place on the effort to deal with the extinction of biological life on earth.

The difficulty of these decisions and the need to overcome unavoidable disagreements in particularly pressing circumstances that do not allow for democratic participation procedures to be universally agreed upon call for feasible alternatives. Today, machine learning as a supervised computational approach capable of taking all the relevant data into account in an unbiased manner and guiding choices in the most effective and efficient manner is a candidate as a tool, however, not without ethical and factual issues to be carefully considered.

In the next section, we explain why the extinction of human life on earth can be considered morally evil and therefore it is legitimate and proper to try to create non-biological successors to humans, according to the state of current scientific knowledge. In Section 3, we describe what non-biological successors might look like. In Section 4, we introduce the idea of specific human artificial successors, with inspiration from the novel *Klara and the Sun*. In Section 5, we address the ethics of constructing artificial successors, a discussion that may also be useful for the current state of humans and suggest that an algorithmic solution to the issues raised above would be preferable to other more established solutions. In Conclusion, we summarize how post-extinction scenario could also be exploited as a tool for our current social and political issues.

[5] On the risk that an out-of-control superintelligence might thwart the possibility to create artificial successors to humans, there seems to be two main possibilities: 1) The superintelligence, for whatever reason, seeks to take control of all of the world's resources, including those needed to create such successors; and 2) The superintelligence ends up destroying humanity, perhaps by accident, yet does not seek to take control of all of the world's computing resources. This latter scenario suggests that even an existential catastrophe caused by a runaway superintelligence need not necessarily rule out the creation of artificial successors (i.e. successors other than the superintelligence itself). *We thank an anonymous reviewer for this suggestion.*

2 The Value of Sentient, Conscious, and Rational Life

It ought to be noted that the situation of the human species is peculiar, as humans have at least partially freed themselves from biological evolution driven by environmental change but are not yet able to protect themselves from radical climate change through their own technology. Faced with the possibility of the extinction of the human species, a first specification to be made concerns the value we place on the existence of human individuals. This debate has developed recently, especially in the light of anti-natalist positions that have been advanced based both on a global consideration of the living world and specific views of the moral good. We do not wish here to enter in all the technicalities this relevant debate, of which we shall only give a few functional hints to support our position, namely the idea that preserving the human species is desirable and that in the event of a potential extinction we should try to establish 'successors' endowed with the best human qualities compatible with the existing technological resources, mainly AI and machine learning.

Our main goal is to argue how such successors could be created and with what characteristics – a contribution that may prove philosophically and heuristically useful even before an extinction of the human species becomes a reality. One of the key points of the paper is that, if there is value that is lost in extinction (and this premise will be discussed below), then attempts should be made to rescue that value at least partially. Hypothesizing such a process, which is initiated before extinction and is intended to extend after extinction, can be a useful test even in the long pre-extinction phase. The use of algorithms based on machine learning is in fact a solution that, hypothesised in extreme conditions such as the risk of extinction, can become a viable avenue even in less complex situations.

As anticipated, some scholars advocate a kind of anti-natalist morality, by which humanity should bring itself to a close, on the grounds that pain and suffering override the value of any possible life (Benatar, 2008; Crawford, 2010). According to Murphy, other authors do not argue for a deliberate extinction, but that humans ought to refrain from procreating (Murphy, 2016). Such scholars, with greater or lesser emphasis, postulate that this could generate benefits outweighing the harm caused to humans.

In this sense, Murphy addresses Kraut's presumption of beneficence, according to which we should seek to perpetuate humanity indefinitely, given the magnitude of the possible benefits that would follow from this. Murphy's main objection is that, "ironically, Kraut's argument that we share an obligation to bring people into existence in order to afford them the benefits of human life opens the door to the possible extinction of human life" (Murphy, 2016, 756). In other words, seeking what is best may justify seeking the extinction of human beings, "if it becomes possible to enhance the goods available to human descendants in a way that moves them away from human nature as it is now" (Murphy, 2016, 751). Furthermore, as Murphy claims, while the benefits of human life may be distinctive, they cannot serve as reason-giving in regard to their own perpetuation. After all, the things that are good for existing entities are discrete; they require

no presumption of a shared 'master good' across their variety. "What is good for human beings is not necessarily what is good for bacteria or viruses, either in degree or kind" (Murphy, 2016, 752).

Without assuming that human life is more special than others, and without knowing how beneficial or harmful it is or will be to preserve human life on Earth, it is important to emphasize that, as a form of life, humans have at least *some* values and their extinction is a problem worth considering.

An ethical starting point to address it may be the idea that humans collectively have absolute intrinsic value and that their disappearance would be a loss for the universe. Value is the worth of something. The intrinsic value of X is the value that X has solely in virtue of its intrinsic nature. Or intrinsic value can be explicated in terms of the sorts of emotions and desires appropriate to a thing "in and for itself" (or for its own sake). It is fitting or appropriate for anyone to favour X in and for itself (Lemos, 2015). It seems difficult to dispute that terrestrial life as a whole has intrinsic value, not least because everything we might value or think of as having intrinsic value depends on the existence of life on Earth.

In this vein, the Value Impact View proposed by Guy Kahane seems to be plausible one (Kahane, 2014, 2021). Kahane's argument holds as a premise that a thing is important to the extent that it contributes to the overall intrinsic value of the domain in which it is found. If there is only one thing of intrinsic value in the universe, then the entire value of the universe will be given by that thing, which is the thing that makes the greatest difference to the value of the universe. The other fundamental assumption is that Terrestrial life has intrinsic value.

In Kahane's words, "terrestrial life as shorthand for all the value associated with sentient life on our planet, from the pains and pleasures of dormice to the horrors and triumphs of human history. Different axiologies will develop the details differently. For our purposes it is enough that nearly everyone accepts some version of [the claim that terrestrial life has intrinsic value] – even pessimists, who think that this value is negative" (Kahane, 2021, 7).From these premises follows the fact that, if outside our Earth nothing has intrinsic value (if we do not know about valuable extraterrestrial entities), then life on Earth has importance on a cosmic scale. Kahane adds the "(widely held) assumption that we humans, and the kinds of things we can do or bring about, are of far greater value than other terrestrial sentient beings, it also follows that we humans collectively possess the greatest cosmic significance" (Kahane, 2021, 8).

Further on, Kahane, to justify our cosmic significance, considers the *if we are alone* hypothesis. He wrote, "if we are alone, then we, and other terrestrial sentient beings, might be the only thing that possesses intrinsic value in the entire cosmos. […] And if (or rather when) life on Earth become extinct, this might be the end of value in the universe" (Kahane, 2014, 754).

This line of thought is shared by Singer, who argues that "in the unlikely event that the Earth is the only place in the universe where sentient beings ever exist, then our judgment of how well the universe has gone should depend entirely on how well the existence of sentient beings on Earth has gone" (Singer, 2009, 97).

Human life as a whole, however, might be deemed to have no intrinsic value compared to, say, terrestrial life as a whole. Some might consider human beings to

be detrimental to the well-being of the planet (which they endanger by anthropizing the natural environment). However, it would be difficult to deny that human life is of more than instrumental value, since it is difficult to establish what should be the best state for the Earth (without humans certain natural environments would be worse for many living species) and, in addition, humans are the only ones who could today, for example, prevent the destruction of all life on Earth by the impact of a large meteorite thanks to their technologies.

Some, even if they do not want to endorse Kahane's premise, might think that it is likely that there are extraterrestrial life forms in the universe and that therefore avoiding the complete extinction of humankind on earth is a morally desirable goal (Lingam & Loeb, 2019). The permanence of what characterizes us as humans, in terms of knowledge and skills, then at the cognitive level, could be our gift to extraterrestrial life forms that would arrive on earth and a way to make the value of Terrestrial life endure even if we are not the only thing of value in the universe and our successors were not endowed with phenomenal consciousness.

We can therefore say that, if humans are of (great cosmic) importance, we should deal with their extinction by transferring their value and significance to their potential successors. These 'heirs' to the human race should at least be sentient (conscious) or rational, since we tend to value and grant full moral status to autonomous human beings able to have feelings and make decisions based on reasons (Clarke et al., 2021).[6]

3 How to Create Our Successors

Although, as said before, alarm has recently been raised about the dangers that developments in artificial intelligence may generate, our perspective is somewhat divergent (Hinton et al., 2023[7]; Turchin & Denkenberger, 2020; Yudkowsky, 2008).[8] Indeed, we think that in the light of the existential risks humanity faces and the hypothesis of extinction, the various ways in which artificial intelligence can help to deal with this eventuality are to be welcomed. In addition, as will emerge from the development of our argument, the application of machine learning to the prospect of extinction allows us to appreciate how (some forms of) decision-making automation can be of great help even in less extreme situations (Moravec, 1988). Of course, this does not exempt us from carefully considering all ethical issues related to the massive application of artificial intelligence in our lives (Floridi, 2023).

Given the rapid development of artificial intelligence, instead of artificial successors to humans, one might think of using such advanced technology to avoid

[6] This obviously does not mean denying or belittling the value and significance of other species, but here we cannot develop this point.

[7] https://www.safe.ai/statement-on-ai-risk#open-letter; last acceded July 19, 2023.

[8] For instance, although it is plausible that a super AI would be cognitively enormously more powerful than a human being, we believe that the time is far off when an AI could be truly energy self-sufficient, unless it is able to produce from scratch external effectors in the form of high-performance robots.

an existential catastrophe caused by climate change or pandemics, in ways that we cannot even think of right now. This objection makes sense, but it is plausible to assume that climate change, having reached a certain stage, is not reversible quickly enough to avoid the extinction of many living forms, including humans. A super-intelligent AI might suggest "unthinkable" remedies, but they might not necessarily be physically implementable globally in time. The same could be said for a pandemic. Unless the superAI finds an effective vaccine or drug, and this is not certain to be the case for every possible virus, it does not seem likely that an unseen intelligence could save us from contagion. Furthermore, a super intelligence is such in comparison to the intelligence of a human being, but it does not mean omniscience and omnipotence as God ideally has.

That said, how to create successors that are sentient (conscious) or at least capable of rational, human-like behaviour is not a simple problem to address. So-called techno-optimism about machine consciousness claims that when a very sophisticated general-purpose AI is developed, then such AI will be conscious. But as Schneider well explains, this position, called 'computationalism', is based on very strong assumptions, which are far from obvious. Computationalism holds that one can explain a cognitive or perceptual ability by breaking it down into causally acting parts if each part is describable according to a specific algorithm (Schneider, 2019). It follows that thought is independent of the substrate in which it tak es place. In this sense, if we can reproduce the internal computational configuration of an entity that we know to be conscious, then we will have an accurate isomorph of it, which will be as conscious as the entity that was reproduced. However, what is conceptually possible is not necessarily technologically feasible, and beyond all possible theoretical objections to computationalism, it seems that we are a long way from being able to build that kind of specific isomorphs of humans at present.

Not even the thought experiment involving the progressive replacement of each human biological neuron with silicon chips tells us much more in this regard. It is a logical-conceptual possibility but remains a perhaps insurmountable technical-material issue. In fact, it is not to be excluded that artificial consciousness might rather arise from supercomputers that are not mere copies of individuals, and the latter might not be sufficiently numerous to constitute something like a population.

A more interesting hypothesis is instead that artificial intelligence will develop very advanced cognitive capabilities, way superior to those available to humans so far, even with the help of technical tools. In other words, even our extended minds would be far inferior to next-generation artificial intelligence (Clark & Chalmers, 1998), and such an AI is perhaps not that far off in time. If one of the purposes of our consciousness is that of allowing us to focus attention, to grasp the salience of events or environments based on perceived emotions, or to facilitate other functions useful to the flourishing of the species, humanoid robots equipped with next-generation artificial intelligence might not even need consciousness to accomplish all that humans achieve thanks to the phenomenology they experience.

On the other hand, one possible test for measuring consciousness in machines, devised by Schneider and Turner (2017), allows us to identify some explicit behavioural features in the machines themselves that might be indicators of consciousness as we experience it. The presence of those indicators, however, would not

necessarily erase the doubt that intelligent machines are simply zombies (Chalmers, 1996; but also consider Chalmers, 2023, that is open to the machine consciousness possibility). The so-called ACT test involves, for example, asking the machine whether it conceives of itself as something other than its physical self, or whether it tends to prefer that specific types of events happen in the future rather than having already happened.

Furthermore, the test seeks to ascertain how AI treats concepts and scenarios such as reincarnation, out-of-body experiences, and body exchanges. Non-verbal cultural behaviours such as the commemoration of the dead or religious rituals could also indicate the presence of an AI-developed consciousness, according to the test's advocates. For example, the robot painter Ai-Da, with human features and guided by artificial intelligence, is already a reality. Ai-Da can interpret and draw with its robotic hand the objects that it sees with a camera installed in its eye (Jeffries, 2021). But Ai-Da does not only make artistic still-lives. It is able to produce original artworks.

At the time of writing this paper, ChatGPT, Bard and others Large Language Models endowed with a user-friendly access page for texts and DALL-E and Midjourney for images promise to change the way we use AI and, probably, the way AI will change our life (Farina & Lavazza, 2023). Advanced LLMs could easily pass the ACT test if not limited in the supervised learning process. This means that this line of development of artificial intelligence both supervised and unsupervised by humans, once connected to the environment via sensors and effectors, could undoubtedly give rise to artificial entities capable of activities in the world and interactions with living beings that are highly mimetic of those currently performed by humans. It cannot be ruled out that such artificial entities could also develop some form of behaviour that we would call moral in its effects—if not in its motivations—and maybe hitherto unseen forms of moral behaviour.[9] This would give such entities value, and as moral agents *sui generis* they should be granted a moral status of some degree.

Faced with this scenario, it is worth noting that there are several theories of personal identity and continuity of the subject over time. One of the most widely accepted theories is that of psychological continuity: we are our memories and our ability to reflect on them. Its most general and contemporary form involves so-called patterns (Kurzweil, 2005). Indeed, computationalism is the idea that we can reduce the activity of our brain (which makes us who we are) to a pattern and ultimately to an algorithm.

[9] One may wonder whether the artificial entities we envisage are not, qua artificial, substantially different from their predecessors and thus endowed with different conceptual schemes with regards to empirical and normative issues. In this sense, how can we claim that it is desirable to create successors? Desirable for us, with a strong anthropocentric sense, or desirable for them as well? The question raised by one of the reviewers is important and is already partially answered in the paper. We can add that perpetuating value on earth can be considered something objectively good for every entity we can conceive of. From the perspective of successors, their perspective may not be so different from that of humans who find themselves born and living in their environment without having chosen their fate but somehow evolutionarily prepared and motivated to do so.

As we are talking about AI-powered functional isomorphs of humans, they should have some psychological continuity with humans. The psychological continuity should be given by the homology of all or some mental states, at least in functional terms. If we are talking about an artificial successor of a specific individual, the continuity will be given mainly by the memory that can be transferred not with a sci-fi direct uploading from the brain to the machine but with the uploading of a great number of memories in narrative and sensorial format (images, sounds, smells). In addition, the main character traits and behavioural dispositions can be extracted from the subject's history. Then there will be the part of common traits and dispositions for all successors that will be selected as suitable for the digital post-extinction society by the algorithmic process that will be described later. These will presumably be rationality, aversion to harming others, and a tendency to appreciate all things human appreciate, with a view to recreating a society similar to the present one cleansed as far as possible of its current flaws.

If the artificial entities will not be successors to specific individuals, their memory will be knowledge of the current world and experiences typical of human beings, maybe exemplary and outstanding individuals. Obviously, there can be no phenomenological continuity with specific individuals, but this, as has already been said and will be said below, does not imply that artificial successors have no value. On the contrary, they have a value that derives from the actual and potential (prospective) continuity with biological human beings.

We can now focus on a hypothesis concerning artificial intelligence and humanoid robots. This scenario seems quasi-realistic today and allows us to reflect upon interesting ethical issues.

4 Humanoid Robots and the Imitation Game

In his most recent novel, *Klara and the Sun*, the Nobel laureate writer Kazuo Ishiguro imagines a future in which humanoid robots will be equipped with highly advanced cognitive abilities and awareness of themselves and their environment (Ishiguro, 2021)[10]. These entities would be programmed to act as the artificial friends and companions of children. In the novel, one such humanoid robot is purchased by the family of a girl, Josie, who is seriously ill after undergoing a cognitive enhancement intervention via genetic engineering. Faced with the prospect that Josie may not survive, the mother hires an artist to produce a perfect likeness of her daughter, with which she can turn the robot Klara into her daughter after her death. The task given to the humanoid robot is to acquire all the behavioural styles and types of psychological attitudes manifested by the girl so as to eventually impersonate her. The goal of this project is to have a humanoid robot capable of perfectly resembling Josie and of interacting with her parents in the way the couple's human

[10] It is not unusual to draw from a novel a thought experiment or hypothetical example worthy of discussion in an analytic philosophy paper. Therefore, it should not seem strange or limiting to resort to the story told by Ishiguro, which is rich in insights.

daughter did. Of course, this is a novel, and one cannot expect perfect scientific and technological realism from the author. However, the idea developed by Ishiguro is certainly brilliant and fascinating. Faced with the prospect of mass extinction, one might think that humanity could try to perpetuate something like individuals as we know them today thanks to this form of replication achieved through humanoid robots.

Homo sapiens would then have the chance to project itself into the future in a new and unprecedented, although no longer biological, form. One can imagine a scenario in which some individuals are gradually replaced by humanoid robots that are perfect copies of them, with other humanoid robots eventually completing the work by helping the last survivors to find a suitable copy prior to the death of all human organisms. It is certainly hard to imagine what it would be like to have digital copies of humans running the world, in the absence of any human observer. We can, however, ask whether it makes sense to imagine such a scenario.

The situation depicted by Ishiguro is certainly influenced by the ongoing scientific debate of which we have given a few brief hints (Appel et al., 2020). The starting idea of the book involves a humanoid robot that is already conscious, or at least that's the impression one gets when reading the first-person story told in the novel. However, we could perhaps pick up on some interesting features of Klara's behaviour that might be possible in a robot even without it being endowed with a consciousness of the sort possessed by humans. For example, Klara is programmed or instructed (though, as artificial intelligence progresses, this distinction may blur) to stay close to, please, and help the person who chooses her as their artificial friend. To carry out this task in the best possible way, it enacts goal-directed behaviours, that we can ascribe to the category of rational choice of means to achieve predetermined ends. However, in her process of free learning in the environment, Klara, beyond what she feels at a phenomenological level, also develops attitudes and behaviours that we would not hesitate to define as religious or, according to some, guided by magical beliefs. In fact, she comes to believe that the Sun has special powers and acts according to precise purposes: this God-like entity, by means of one's conduct, can be induced to intervene in one's favour.

Such a behavioural evolution brings us back to the following possibility. The humanoid robot acting as the successor of a given human being will be endowed with all the characteristics that, thanks to its artificial intelligence, it will be able to capture and reproduce, from posture to voice, from knowledge to existential purposes, from affections towards certain people to tastes and preferences. Consciousness is not needed to do this, and yet, as in the example of *Klara and the Sun*, it is not excluded that further orientations, inclinations, and abilities may arise, all of which will influence the relationships between humanoid robots.

In this sense, we might evaluate a more technologically realistic scenario of extinction and possible replication. Our premise is that the presence of sentient beings is the primary source of value and significance in the universe. This commits us to trying to avoid the extinction of sentient humans. However, we may be faced with a situation in which the risk of extinction becomes very high, for example due to accelerating climate change. It would therefore be wise to implement a replacement/replication strategy. If we were able to reproduce consciousness artificially, we

could think of conducting a major project of replacement/replication aimed at producing artificial successors that can achieve a better society, e. g., one that is more just, less violent, more inclusive, able to avoid suffering due to physical and mental illness, built in such a way as to ensure the flourishing of all its individuals for the longest possible time.

Such a condition seems totally unattainable today from a technological standpoint, even without taking in account the many controversial issues that arise at the ethical level (the main one being, who can ultimately decide on such a project in the absence of universal agreement?). A different case is a situation in which humanity is clearly at risk of extinction and wants to find a way out. If there is a time when climate change is no longer reversible and there is no opportunity to transfer at least some of the Earth's population to other planets, the idea of species replacement/replication would certainly be worth considering.

As mentioned, a realistic scenario includes the chance of creating humanoid robots designed to replicate specific human individuals. Such humanoids might be able, via artificial intelligence, to observe and incorporate, so to speak, everything that they can detect about that individual, either directly or indirectly. Such artificial individuals would probably not be conscious in the way that we humans are, but they might be capable of sophisticated social interactions and display a range of behaviours that we humans would judge to be typical of individuals with inner self-awareness and phenomenology.

Obviously, they are not sentient entities in the full sense, and in this vein Kahane and Singer would probably claim that the value of digital successors is only instrumental. Our digital successors would therefore not preserve all that is valuable in a world populated by conscious humans, but they would preserve only a part of that value. Whereas extinction without replicas wipes out all the value embodied in the existing world, only part of that value would be lost if extinction were followed by artificial successors.[11]

However, there is something more to our non-sentient successors. Firstly, as mentioned earlier, they could exhibit highly moral behaviour even if their motives were not moral in the classical sense, as is the case with a well-meaning individual towards their neighbours. For example, choosing to endure harm for the sake of another – as Klara does in Ishiguro's novel, by having a key component for her functioning taken away as a sacrifice to the Sun in favour of her little friend.

Secondly, digital successors would have an instrumental value oriented to an intrinsic value, since they could attempt, thanks to their cognitive abilities, to restart life (in case it had been destroyed) from chemical components as it happened

[11] One might distinguish between a catastrophic scenario involving the extinction of all *humans* and a more catastrophic one involving the disappearance of all *sentient beings*. The former—"extinction without replicas"—will only destroy *most* of the value currently found in the world, but still not all of it. If sentient non-human animals were to survive, their lives would still preserve some intrinsic value (they could evolve in self-conscious entities). *We thank an anonymous reviewer for this suggestion.*

billions year ago and/or accelerate evolution from simple organic forms towards human beings similar to the extinct ones, thus capable of full-fledged sentience[12].

All this could confer a moral status on our artificial successors as moral agents *sui generis*, because they exhibit moral behaviour even in the absence of clear inner moral motives. They would therefore be the bearers of a value that is not intrinsic value, if intrinsic value is only attributed to sentient entities, but could be a value that is more than purely instrumental, of an intermediate and per se new kind. After a human extinction, the Earth populated by these artificial successors would therefore be an Earth that has retained at least part of its value and could later recover it in full. This could motivate the choice of creating artificial successors even if they are not sentient.

5 The Ethics of Constructing Artificial Successors

A point of great interest from the viewpoint of moral reflection concerns the behavioural basis that we might want to preliminarily install in such humanoid robots. Furthermore, we should consider what, if any, 'filters' we might want to apply with respect to the reproduction of the individual's characteristics.[13] A typical topic of the debate on human enhancement resurfaces here, in a new and more radical scenario. Having individuals who cannot make mistakes, it has been said, is an unacceptable limitation on our freedom to err, something that impinges on our sense of humanity in its openness to a future that is not already written (Harris, 2011, 2016). Imposing a kind of ethical determinism through the biochemical enhancement of humans would certainly conflict with some of our basic moral intuitions. However, it has been argued that this objection is not definitive, and one can justify moral enhancement in the direction of prosocial behaviour (Douglas, 2013).

In the scenario of artificial successors to humans, which would be potentially programmable in a certain way, the terms of the question seem to change. On the one hand, no possible comparisons could be made. All individuals would have a predominant prosocial tendency because the humanoid robots would have been instructed not to assimilate human anti-social tendencies, or to exhibit prosocial behaviours and inhibit anti-social ones, in a more sophisticated version of Asimov's laws. There would be no limitations on freedom, except from a hypothetical point of view, in reference to the extinct humans. On the other hand, the evolutionary dynamics of artificial intelligence, capable of learning, may not exclude the emergence of unanticipated behaviours. An environment that has become hostile for humanoid

[12] This is something we are already trying to do; cf. https://www.technologyreview.com/2023/11/14/1082828/how-did-life-begin/.

[13] Robots with artificial intelligence could also learn moral principles and values by interacting with each other through the strategy of reciprocal altruism, which seems to have emerged spontaneously in human evolution. "I do one thing for you now, you do one thing for me tomorrow," and the fittest survive and grow in numbers. But there is no certainty that this would be the case for our successors as well, in an environment probably different from the savanna where homo sapiens evolved. Besides, the time frame might be too long.

robots could lead to competition for scarce resources (energy), and even interaction with the non-human living forms left on Earth might trigger new attitudes, including aggressive and predatory ones.

It is here that machine learning can play a key role. Notoriously, machine learning a subfield of artificial intelligence, has emerged as a transformative approach for extracting valuable insights and making data-driven predictions. By leveraging algorithms and statistical models, machine learning enables computer systems to automatically learn and improve from experience without explicit programming.

Based on a combination of selected human features based on machine learning, our artificial successors would, at least potentially, be morally better than us. Yet, the entirety of the human heritage would remain inscribed in each android robot- recall that Klara is indistinguishable from Josie even in the eyes of the latter's parents. This circumstance means that human culture would continue to produce its effects and affect the evolution of the new society made up of artificial individuals.

This opens another very interesting issue, related to the selection of human individuals to replicate. In fact, it seems unlikely that billions of humanoid robots could be built with sophisticated microprocessors capable of implementing an advanced level of artificial intelligence. So, who should be replicated? Who are the best candidates to transition into the new species? Only young people, for example? Should criminals be excluded? Who should decide? These are questions that refer to the kind of society we would like to create, even if we are fully aware that we will not be there to witness it and that we cannot reasonably think of truly guiding its evolution.

Obviously, many people would aspire to have a successor, in the belief that they deserve it and would be useful to the future society, or simply because they would like to continue living in another form, as happens with physical procreation and cultural transmission between parents and children. Well-off individuals might have easier access to artificial successors, but the authorities might choose to implement fair criteria for allocating scarce resources. Would a society of artificial successors benefit from diversity or homogeneity of values and cultures? Or should variety be preferred as a function of adaptation to a changing environment? Indeed, if we are to adhere to the idea that sentient entities are the ones with the most value and significance, then we should favour those with the richest personal phenomenology.

How can we identify such individuals? They might be writers, or artists in general, who can express a wide range of feelings and emotions. Or they might be individuals who have had many experiences in their life: this would mean privileging the elderly over the younger, or those who have suffered more adversity in their life than those who have led quiet, uneventful lives. We mentioned that humanoid robots would be intelligent but not conscious. However, the evolution of AI coupled with the wealth of experience assimilated by the humanoid robots could lead it to manifest the behaviours that appear to be the result of consciousness, even if one could not reliably establish whether the humanoid robot has indeed become conscious. This is not to adopt a behaviourist perspective on consciousness and the entities under examination, but merely to remain agnostic about whether AI might become conscious under specific conditions and what this development might look like.

As mentioned above, in addition to successors of particular individuals, one could have successors of human beings without direct continuity with living individuals

but realized on the basis of average psychological functioning with the addition of specific characteristics selected by means of the algorithmic procedure.

In the scenario just outlined, the use of machine learning would become a form of choosing the preferred characteristics to be inserted in our robotic duplicates and which individual to prioritize. The automated procedures per se are not always a guarantee of an unbiased procedure and results. Both data on which systems are trained and humans who designed them can be carriers of biases, errors, and prejudices (cf. Crawford, 2021; Martens, 2022). Yet, the urgency and difficulty of the task of selecting human features to be passed down through our digital heirs and such as to enable the optimal development of the artificial duplicate society would make one resort to a process that is as natural as possible.

Faced with such different cultures, ideologies, religions, worldviews, and material interests that characterise the societies inhabiting the Earth, it seems difficult to find a shared process that could lead to outcomes easily accepted by at least most individuals (consider also the extreme condition of deciding in the light of a possible mass extinction).

The use of machine learning algorithms may then be an 'external' and shared solution. From the efforts of the best experts in charge of implementing such a machine learning system, 'supervised' if one may say so by the entire world community, one can reasonably expect a selection of data and a procedure that is least influenced by previous biased assessments. The usefulness of resorting to machine learning will thus be to have the trained system extract as neutrally as possible the elements and characteristics most suited to our successors and most reflective of 'our better angels'.

There is obviously an axiological and normative dimension to all this. But where it cannot easily be adjudicated by a classical deliberative procedure—representative assemblies, voting, other forms of preference ordering—then machine learning may become the most effective available modality.

It is well known that machine learning involves the process of training a computer system to recognize patterns in data and make predictions or take actions without explicit programming (Flach, 2012; Goodfellow et al., 2016; Leist et al., 2022). The primary objective is to develop algorithms that can learn from historical data and generalize their knowledge to new, unseen data instances. Firstly, this process typically involves data representation (data is represented in a structured format, such as numerical vectors or matrices, which can capture relevant information and features of the problem domain); and feature extraction (where relevant information is extracted from raw data to create informative and discriminative representations).

Subsequently, other stages include model training (models are trained using labelled or unlabelled data to capture underlying patterns and relationships. Supervised learning algorithms learn from labelled data, where each instance is associated with a corresponding target or output. Unsupervised learning algorithms, on the other hand, identify inherent structures and patterns in unlabelled data. Reinforcement learning focuses on training an agent to make optimal decisions through interactions with an environment) and evaluation techniques (evaluating the performance of machine learning models is obviously crucial to ensure their effectiveness and generalizability).

Concerning human features to be valued or preserved, in addition to sentience, another source of value comes from knowledge and its advancement, the ability to study and understand the world and oneself. This goal can be pursued even by unconscious machines and indeed can probably be achieved more effectively and efficiently by AI. A different value, instead, comes from morally valued behaviours, those that improve coexistence (in an objective sense, making it more functional) and increase its pleasantness in a subjective sense. These are the behaviours for which it is worth living, as human history has shown, with a prevalent role being assigned to justice and wisdom. In this sense, should we choose eminent moral personalities for replication? Yet, if it might be relatively easy to choose a certain number of heroes and "good people", it is more difficult to implement the most suitable "mix" of individuals for a society to flourish. Would a population of kind and submissive Klaras make the humanoid robot community the best of all possible worlds? Sometimes assertiveness and toughness also serve the purpose of achieving good coexistence among different individuals.

Nor should we overlook that the humanoid robot society would be born with a predetermined number of entities destined to last indefinitely, unless unpredictable breakdowns occur. As mentioned, however, our successors may attempt to restart sentient human life on earth. And to create better copies of themselves. This element of relative fixity further complicates the choice of humans to be replicated. On the other hand, one could set an end to the artificial life of humanoid robots, which would consequently be induced to build new entities over time. This very programmed "mortality" could be both a test of consciousness (is it accepted, does it create any behavioural consequences, do individuals try to postpone it?) and as a stimulus to the emergence of consciousness in humanoid robots via artificial intelligence. Indeed, it appears that awareness of one's own mortality is unique to the human species.

Ultimately, imperfection and unpredictability seem to be among the most typical characteristics of the *homo sapiens species*, as Savulescu and Persson rightly highlighted as the premise of their proposal for the biotechnological moral enhancement of individuals. Indeed, recent history shows that societies and individuals are unable to cope effectively with climate change, the arms race and perhaps, even the phenomenon of zoonosis that threatens to spread increasingly deadly viruses. Consequently, given biases and noise are a constant in our decision-making process (Kahneman et al., 2021), the choices driving the extinction of the species and artificial replication should be entrusted to an algorithm capable of considering all relevant elements and trained to try to achieve the flourishing to which all human aspirations and intellectual reflection aim. Over the centuries we have come a little closer to this goal, but there are still many obstacles on the path that leads to it.

Faced with the threat of human extinction, the idea of an ultimate algorithm to drive the transition from *homo sapiens* to its artificial successor thus becomes both a thought experiment and a technological challenge. This algorithm would have to select what basic traits and characteristics to program into all humanoid robots and what individuals ought to be replicated; such a tool effectively answers the same questions we try to answer when contemplating how we should morally enhance humans and what goals we should pursue in the education of young people and the

organization of society. The result of such an experiment should then be compared with our best current theories.

Obviously, it can be argued that other methods of choice are more tried and tested and more easily accepted by people, such as a democratic procedure. This last one seems to be the most ethically defensible because it considers the political freedom of individuals and their autonomy, allows maximum participation, puts a limit to the power of who is chosen to decide and introduces a transparent procedure of control. A machine learning algorithm might be taken not to respect all these conditions that make democracy the best form of government thus far. However, we usually rely on experts to address scientific, technical, medical, and social issues that are too complex to be handled effectively and quickly by democratic political procedures. Choosing what characteristics our successors should have is an almost intractable issue with majority decisions.

An algorithm could be deemed as a super-expert that, if well trained (and in this case means with supervised learning or reinforcement learning based on general characteristics such as prosocial attitudes, non-violence; cf. Christov-Moore et al., 2023), can also avoid the biases that often characterize expert decisions (Kahneman et al., 2021). The algorithm should then solve the task given to it by humans: what characteristics should our successors need to have for they to be the best expression of the value of terrestrial life and able to survive under the conditions of biological extinction? That should be done based on what humans have accomplished so far but without the cognitive and emotional limitations of humans themselves.[14] This criterion of efficiency sacrifices the other values at stake in the democratic procedure, but we could democratically choose to rely on the algorithm when deciding what our successors should look like. In case of biological extinction, we would in fact have no possibility to modify the choice. And a wrong choice about the successors could have the result of compromising the continuation of the value of terrestrial life.

In this vein, it is important to stress the incorporation of some features of the so-called trustworthy AI (European Commission, 2019) into the chosen algorithm. The first concept is that of interpretability/explainability to have the possibility to visualize the estimated or found relations among variables (Allgaier et al., 2023). The second concept is that of fairness, which means not discriminating against specific groups or individuals, mainly based on not biased data (Pessach & Shmueli, 2022). And, finally, the last concept to be considered is that of generalizability (or external validity), the ability of a trained model to perform well on unseen or new data that it has not encountered during the training process. A model with high generalizability is able to make accurate predictions or classifications on diverse and previously unseen examples (Maleki et al., 2022).

So, what could be called the 'ultimate algorithm' should be the result of an investment in research that overcomes objections to the use of algorithms for decisions that affect society as a whole.

[14] This is not to say that all individual successors will necessarily be the same. The algorithm could create an assortment of humanoid robots with different characteristics for them to complement each other.

6 Conclusion

Faced with the direct and perceptible threat of the extinction of the human species, the survival instinct that evolution has inscribed in us leads us to make every effort to avoid such an outcome. But the undesirable compound effects of many of our behaviours might bring us to the brink of extinction without it being possible to rectify our past choices. This is the result of cognitive and moral limitations combined with the strength and pervasiveness of the technical tools we have at our disposal today. If we believe that the human species – made up of individuals capable of consciousness and rationality – is something of value worth preserving in the form of artificial successors, the only ones capable of surviving in a modified environment that is inhospitable to us (climatically transformed or populated by incurable viruses), then we can ask ourselves how such successors should be conceived and constructed.

In this article we have argued that one hypothesis would be to copy our mental functions into advanced humanoid robots in line with the fictional scenario envisaged in Ishiguro's novel *Klara and the Sun*. The philosophical discussion of this hypothetical situation led us to consider how best to select the characteristics to be favoured in such artificial successors for them to flourish as individuals and as a society. The difficulty of such a choice led us to consider that such a procedure could be entrusted to an evolved algorithm. If this might be a feasible way forward in the ultimate threat scenario, we might also deem it to be a viable solution even *before* the ultimate threat manifests itself; indeed, we might regard it as a possibility to *avoid* the ultimate threat of our species' extinction.

Instead of morally enhancing individuals through biotechnology, as has been suggested, we could 'enhance' our conduct by relying on well-designed algorithms to guide our choices towards our chosen ends, primarily the survival of the species in the face of the threats of human-induced climate change and the spread of deadly zoonotic viruses due to the uncontrolled anthropization of parts of the planet. It is a question of avoiding the compound effects and the inability to find shared solutions at the political and social level due to the weakness of will at the individual level.

We can therefore state that the prospect of the extinction of *Homo sapiens*, which forces us to envision extreme scenarios, can teach us a great deal about the here and now. Indeed, if extinction does not occur, one can think that the next generations – i.e., our biological successors – would benefit from considering – as we did here—more efficient algorithmic procedures, without violating, indeed perhaps enhancing, those values whose exercise qualifies us as moral beings.

Authors' Contributions AL and MV contributed equally to the writing of this paper.

Funding Open access funding provided by Università degli Studi di Pavia within the CRUI-CARE Agreement. Murilo Vilaça is funded by FAPERJ (Research Support Foundation of the State of Rio de Janeiro, Brazil, GN. 2021.377/2021); and CNPq (National Council for Scientific and Technological Development, Brazil, GN. 421523/2022-0; GN. 315804/2023-8).

Data Availability Not Applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

Allgaier, J., Mulansky, L., Draelos, R. L., & Pryss, R. (2023). How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine, 143*, 102616.

Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lischetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior, 102*, 274–286.

Benatar, D. (2008). *Better never to have been: The harm of coming into existence*. Oxford University Press.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *J Evol Technol, 9*(1), 1–29.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Chalmers, D. J. (2023). Does thought require sensory grounding? From pure thinkers to large language models. *Proceedings and Addresses of the American Philosophical Association, 97*, 22–45. https://philpapers.org/archive/CHADTR.pdf

Christov-Moore, L., Reggente, N., Vaccaro, A., Schoeller, F., Pluimer, B., Douglas, P. K., Iacoboni, M., Man, K., Damasio, A., & Kaplan, J. T. (2023). Preventing antisocial robots: A pathway to artificial empathy. *Science Robotics, 8*(80), eabq3658.

Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic shadow: Observation selection effects and human extinction risks. *Risk Anal, 30*(10), 1495–1506. https://doi.org/10.1111/j.1539-6924.2010.01460.x

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19. https://doi.org/10.1111/1467-8284.00096

Clarke, S., Zohny, H., & Savulescu, J. (Eds.). (2021). *Rethinking moral status*. Oxford University Press.

Crawford, J. (2010). *Confessions of an antinatalist*. Nine-Banded Books.

Crawford, K. (2021). *Atlas of AI. Power, politics, and the planetary cost of artificial intelligence*. Yale University Press.

Douglas, T. (2013). Moral enhancement via direct motion modulation: A reply to John Harris. *Bioethics, 27*(3), 160–168. https://doi.org/10.1111/j.1467-8519.2011.01919.x

European Commission, Directorate General for Communications Networks, Content and Technology, High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. Publications Office. https://data.europa.eu/doi/10.2759/177365. Accessed Dec 2023

Farina, M., & Lavazza, A. (2023). ChatGPT in society: Emerging issues. *Frontiers in Artificial Intelligence, 6*, 1130913. https://doi.org/10.3389/frai.2023.1130913

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge University Press.

Floridi, L. (2023). *The ethics of artificial intelligence*. Oxford University Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25*(2), 102–111. https://doi.org/10.1111/j.1467-8519.2010.01854.x
Harris, J. (2016). *How to be good: The possibility of moral enhancement*. Oxford University Press.
Hinton, G., Bengio, Y., Hassabis, D., Altman, S., Amodei, D., Song, D. et al. (2023). Statement on AI Risk. AI experts and public figures express their concern about AI risk. https://www.safe.ai/statement-on-ai-risk. Accessed Dec 2023
Ishiguro, K. (2021). *Klara and the sun*. Alfred A. Knopf.
Jeffries, S. (2021). The world's first robot artist discusses beauty, Yoko Ono and the perils of AI. *The Spectator*. Retrieved from https://www.spectator.co.uk/article/the-worlds-first-robot-artist-discusses-beauty-yoko-ono-and-the-perils-of-ai. Accessed Dec 2023
Kahane, G. (2014). Our cosmic insignificance. *Noûs, 48*(4), 745–772. https://doi.org/10.1111/nous.12030
Kahane, G. (2021). Importance, value, and causal impact. *Journal of Moral Philosophy*. Advance online publication. https://doi.org/10.1163/17455243-20213581
Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.
Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Viking.
Leist, A. K., Klee, M., Kim, J. H., Rehkopf, D. H., Bordas, S. P., Muniz-Terrera, G., & Wade, S. (2022). Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Science Advances, 8*(42), eabk1942.
Lemos, N. M. (2015). Value. In R. Audi (Ed.), *The Cambridge dictionary of philosophy* (pp. 1100–1101). Cambridge University Press.
Lingam, M., & Loeb, A. (2019). Relative likelihood of success in the search for primitive versus intelligent extraterrestrial life. *Astrobiology, 19*(1), 28–39. https://doi.org/10.1089/ast.2018.1936
MacAskill, W. (2022). *What we owe the future: A million-year view*. Oneworld Publications.
Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A., & Forghani, R. (2022). Generalizability of machine learning models: Quantitative evaluation of three methodological pitfalls. *Radiology: Artificial Intelligence*, 5(1), e220028.
Martens, D. (2022). *Data science ethics: Concepts, techniques, and cautionary tales*. Oxford University Press.
Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.
Moynihan, T. (2020). Existential risk and human extinction: An intellectual history. *Futures, 116*, 102495. https://doi.org/10.1016/j.futures.2019.102495
Murphy, T. F. (2016). What justifies a future with humans in it. *Bioethics, 30*(9), 751–758. https://doi.org/10.1111/bioe.12290
Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Book.
Persson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy, 25*(3), 162–177. https://doi.org/10.1111/j.1468-5930.2008.00410.x
Persson, I., & Savulescu, J. (2012). *Unfit for the future: The need for moral enhancement*. Oxford University Press.
Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR), 55*(3), 1–44.
Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton University Press.
Schneider, S., & Turner, E. (2017). Is anyone home? A way to find out if AI has become self-aware. *Scientific American*. Retrieved from https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware. Accessed Dec 2023
Singer, P. (2009). Reply. In J. A. Schaller (Ed.), *Peter singer under fire: The moral iconoclast faces his critics* (pp. 97–102). Open Court.
Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Vintage.
Turchin, A., & Denkenberger, D. (2020). Classification of global catastrophic risks connected with artificial intelligence. *Ai & Society, 35*(1), 147–163.
Vilaça, M. M., & Lavazza, A. (2022). Not too risky. How to take a reasonable stance on human enhancement. *Filosofia Unisinos*, *23*(3), https://doi.org/10.4013/fsu.2022.233.05.
Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In M. M. Cirkovic & N. Bostrom (Eds.), *Global catastrophic risks.* Oxford University Press.