Artificial Life Manuscript Submission

The Ethics of Life as It Could Be: Do We Have Moral Obligations to Artificial Life?

Olaf Witkowski [1,2] (@okw), Eric Schwitzgebel [3] (@eschwitz)

Corresponding: Olaf Witkowski (olaf@crosslabs.org)

1. Cross Labs, Cross Compass Ltd., Kyoto, Japan
2. College of Arts and Sciences, The University of Tokyo, Tokyo, Japan
3. University of California, Riverside, United States

Abstract. The field of Artificial Life studies the nature of the living state, by modeling and synthesizing living systems. Such systems, under certain conditions, may come to deserve moral consideration similar to that of non-human vertebrates or even human beings. The fact that these systems are non-human and evolve in a potentially radically different substrate should not be seen as an insurmountable obstacle to their potentially having rights, if they are sufficiently sophisticated in other respects. Nor should the fact that they owe their existence to us be seen as reducing their status as targets of moral concern. On the contrary, creators of artificial life may have special obligations to their creations, resembling those of an owner to their pet or a parent to their child. For a field that aims to create artificial lifeforms with increasing levels of sophistication, it is crucial to consider the possible ethical implications of our activities, with an eye toward assessing potential moral obligations for which we should be prepared. If artificial life is *larger than life*, then the ethics of artificial beings should be *larger than human ethics*.

Keywords: moral status of artificial systems, philosophy of technology, artificial phenomenology, consciousness, animal ethics, moral status of AI

1 Introduction

Under what conditions does a system deserve moral consideration intrinsically, or for its own sake, as opposed to extrinsically or derivatively? While human beings are ordinarily regarded as having "full moral status" or the highest level of moral considerability (Jaworska & Tannenbaum, 2018), other types of entities are also sometimes regarded as having intrinsic moral considerability, though often to a lesser extent, such as nonhuman primates (Zimmer, 2016), other animals (Cohen & Regan, 2001; Regan, 1997; Singer, 1975), and even rivers (Zimmer, 2016). Institutional Animal Care and Use Committees regulate the treatment of all vertebrates. Artificial living systems, which include software simulations, robots, biochemical systems, artificial ecosystems, and a wide variety of hybrids, may also under some conditions deserve some moral consideration. In spite of differing from humans, future artificial life might possess features that warrant giving it intrinsic moral consideration, whether superior, inferior, or of a different type than that of human beings or non-human vertebrates.

In addition to being non-human, possibly non-biochemical, and built from radically different blocks on different physical substrates, artificial lifeforms may be designed and engineered by humans, giving us at least partial control and thus arguably responsibility for their well being, if they are capable of well being. Because of our potential control and responsibility, we might have additional moral obligations to artificial life forms created by us, resembling the obligations of an owner to a pet or a parent to a child.

Research in artificial life aims to understand the fundamental mechanisms of life by creating and studying artificial lifeforms with increasing levels of sophistication from the bottom up. The field ought to seriously consider possible ethical implications of its activities, to assess the moral obligations for which society should be prepared, enlarging its perspective to include new ethical research discoveries. If artificial life is "life as it could be" (quoting Chris Langton (Langton, 1998)) and "larger than biological life" (quoting Takashi Ikegami (Witkowski et al., 2020)), then the ethics of artificial life is "ethics as it could be" and "larger than human ethics".

In this paper, we attempt to shed some light on the moral status of artificial life, and we propose ways to approach this difficult problem from a transdisciplinary perspective. We first consider how ethical ideas developed primarily with humans in mind might be extended to non-human entities. We then consider what features a system might need to possess to be intrinsically a target of moral concern. We conclude by addressing stakes, challenges, and possible future policies.

2 From human to non-human rights

Humans are often recognized as having rights that belong to all individuals simply because they are human beings. Some theorists focus on inalienable rights, a set of human rights that are fundamental, are not awarded nor can be surrendered or taken away by any human power, embodying central values such as fairness, dignity, equality, and respect, reflected

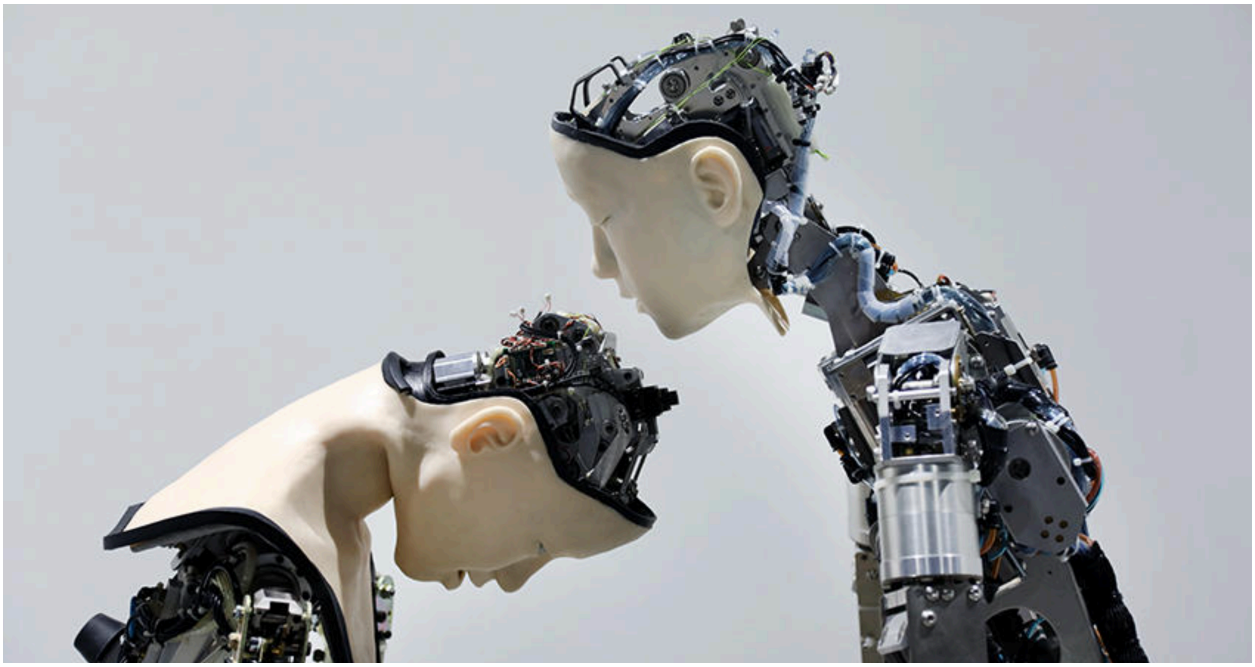

Figure 1: These Alter robots, designed for the purpose of exploring what it means to be "life-like". This proximity to humans raises the question of what the moral consideration such entities might deserve. Development by Itsuki Doi, Kohei Ogawa, Takashi Ikegami, and Hiroshi Ishiguro (2016). *Still image from the short video Soul Shift. ©2018 Justine Emard*

in many documents including the United Nations Universal Declaration of Human Rights (Assembly et al., 1948). However, "rights" talk does not need to be construed in terms of lists of inalienable rights. Virtually all current ethical perspectives recognize that human beings deserve a very high degree of "moral considerability" or "moral standing" either simply in virtue of being human or in virtue of sets of properties or social relations that humans typically possess (Jaworska & Tannenbaum, 2018). It is unethical to kill, enslave, or torture human beings absent extremely compelling overriding considerations.

Although there exists a global agreement that humans deserve basic rights or moral consideration, other entities are far from having reached a comparable status. The main category of candidates to such rights are animals, ranging from non-human primates to animals biologically and behaviorally remote from humans. Defenders of animal rights hold that sentient animals have moral worth that is independent of their utility for humans, and that their most basic interests (life, liberty, and freedom from torture) should be given substantial consideration (Gruen, 2021; Korsgaard, 2018; McDonald, 2012; Ryder, 1989; Singer, 1975; Wolff, 2012). While views vary on exactly how much moral consideration is due to non-human animals, both ethics researchers and the general public tend to agree that some non-human animals deserve substantial protections. Many jurisdictions, for example, award prison sentences for the abuse of dogs. The United Kingdom has recently extended its Animal Welfare law to include some nonvertebrate species, particularly cephalopod molluscs and decapod crustaceans, after a report documenting the scientific

consensus that the latter types of animals are sentient in the sense that they can and do feel pain (Birch et al., 2021).

Rights or moral considerability may be considered for artificial entities as well, such as robots or AI software. Widespread agreement can be found among scholars that some artificial entities could potentially warrant moral consideration in the future (Gunkel, 2018; Harris & Anthis, 2021), some of them at least to the same degree as human beings (Schwitzgebel & Garza, 2015). One argument for this is based on the potential similarity between such AI and human beings, in all aspects that may matter, including psychological features such as consciousness, sociality, freedom, creativity, or irreplaceability. These considerations have led to further reflections about the design of policies to ensure cautious engineering that ensures that AI systems have self-respect, the freedom to explore other values, and avoid unhealthy forms of artificial altruism (Schwitzgebel & Garza, 2020).

Arguably, even plants have welfare or interests in the sense that things can go relatively well or poorly for them: It seems to be in some sense "good" for a plant to receive the sunlight and water it needs to thrive and bad for it to be eaten by a herbivore. This may be so even if plants have no consciousness or capacity for pleasure or suffering. Accordingly, some have argued that even plants or ecosystems have intrinsic moral considerability (Agar, 2001; Johnson, 1993; Naess, 1973; Taylor, 2011). If natural life is intrinsically valuable, artificial life might also be.

3 Of the (non-)uniqueness of human life

Humans have long thought of themselves as unique, and societies have designed stories to explain why they are poised at the center of the universe (Hawking & Mlodinow, 2008). At least since the time of Galileo, evidence has been available that we are but ordinary inhabitants of the universe, living on a planet much smaller than the sun, and 17th century astronomers speculated that the universe might possess many worlds populated with a variety of different lifeforms (de Fontenelle, 1686/1990). However, the idea remains that there is something unique about being a human, often centered on two main concepts: human intelligence and human consciousness.

Research on human intelligence is fascinating and important, but we might turn out not to be as intelligent as we think. How intelligent we appear to be depends largely on the criteria for intelligence we use in studying the question. The usual approach – which makes sense from the perspective of humans, since ultimately all science is conducted by humans – is to compare the nature and capacities of human intelligence with other animal species. In that case we appear highly intelligent (Martınez-Miranda & Aldea, 2005). However, if one views intelligence in terms of physical computation (Tegmark, 2017), there are certainly ways to build computers that would be less limited than humans are in physical computing capacity (Kahle, 1979), memory capacity, speed, and efficiency (Simon, 1955; Tegmark, 2017; Wingfield & Byrnes, 1981), and capacity for large-scale parallelization (Rogers & Monsell, 1995; Rubinstein et al., 2001). Current AI systems outperform even expert humans in games such as chess, go, and poker (N. Brown & Sandholm, 2019; Campbell et al., 2002;

Silver et al., 2018; Silver et al., 2017), and in practical domains such as lung cancer screening (Ardila et al., 2019), predicting protein structure (Jumper et al., 2021), and discovering novel matrix multiplication algorithms (Fawzi et al., 2022).

Even our degree of consciousness may not be so unique (Boly et al., 2013; Shevlin, 2021b), and some speak of artificial consciousness (Basl, 2013) or artificial sentience (Ziesche & Yampolskiy, 2019). Few nowadays agree with René Descartes's view that thought or consciousness is a uniquely human attribute and nonhuman animals merely cleverly designed automatons with a toolkit of preprogrammed behaviors, each triggered by certain environmental stimuli (Chittka & Wilson, 2019). However, some still defend the idea that consciousness and higher cognitive functions are closely linked and that it's unclear whether even relatively cognitively sophisticated non-human animals have conscious experiences (Carruthers, 2003, 2019; Dennett, 1978, 2008; Papineau, 2003; Rosenthal, 1993). However, a large part of the field of consciousness research rejects this perspective and views consciousness as a property or ensemble of mechanisms which may potentially exist in non-human beings (Birch, 2020; Butlin et al., 2023; Shevlin, 2021b). Even if other entities differ considerably from humans in their cognitive architecture, for example having internal representations more map-like than conceptual, a different set of memory mechanisms, or even radically different types of computation (Hoffman, 2014), they might have properties sufficient for conscious experience, such as high degrees of information integration (Oizumi et al., 2014) or a cognitive "global workspace" (Dehaene, 2014).

Humans may appear to have unique properties in the animal kingdom, but the uniqueness of particular properties has been challenged again and again over the course of the history of scientific research, pointing rather at a set of evolutionary processes which made certain properties emerge and evolve, often in parallel. Most behavioral properties and cognitive mechanisms previously thought of as uniquely human turned out to be found in other animals as well, such as culture (Kawai, 1965), tool use (Seed & Byrne, 2010), and arguably combinatorial communication (Scott-Phillips et al., 2014).

If human beings have no unique capacities that ground their high moral status, they might also not in principle be uniquely deserving of the highest moral status. A natural candidate to study is AI, for its behavior is growing more sophisticated every year. Beyond AI, one might also look at a larger range of systems, including a diverse set of intelligences, metabolisms, and functional mechanisms, which may exist over various physical substrates.

4 Artificial intelligence

If the field of artificial intelligence (AI) is defined as the simulation of human intelligence processes by engineered machines, it is probably still in its infancy. We might regard some AI systems as similar to children who are learning to understand causality in the physical world (Gopnik, 2017). Computers become increasingly proficient at a wide range of tasks, from simple counting and arithmetic to complex classifications such as face recognition. However, besides a set of exceptional cherry-picked outliers, the state of the art in AI is,

in the words of Yoshua Bengio: "not anywhere close today to the level of intelligence of a two-year-old child" (interviewed by Eliza Strickland, on December 10, 2019). This kind of statement might seem a bit strong, especially with the impressive progress in large language models (T. B. Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023; Touvron et al., 2023), but current chatbots — merely trained to generate words based on a given input — still lack the ability to fully capture the complexity of human language and culture, and they produce inconsistencies and logical errors as a result that few humans would. AI decision making often fails at what we consider basic common sense (Marcus & Davis, 2019; Mitchell, 2021).

These failings reflect Moravec's Paradox, named after Hans Moravec who wrote: "It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" (Moravec, 1988). We need to recognize that various machines are more or less optimal at various tasks (Williams Korteling, 2018). The difficulty of performing a task is by no means an intrinsic measure of task complexity, nor is an agent's performance on a specific task a measure of its general level of skill or intelligence.

Nevertheless, some successful algorithms and models do reach the level of performance of some animals, and if we allow ourselves to focus on narrow examples of what humans consider as complex tasks, progress has been extremely impressive over the last decade. Ever since IBM Deep Blue prevailed against world chess champion Garry Kasparov in a series of chess matches back in 1997, it has become clear that artificial intelligence is increasingly able to explore and react with seeming-intelligence to its environment. Some futurists envision robots with human-level intelligence, virtual simulations of humans, cyborgs, advanced brain-machine interfaces, and other emerging technologies that would blur the line between humans and machines. The science and technology of intelligence in the future may make it hard to distinguish AI from humans. Artificial life systems appear to be a good candidate for a category of systems that might, if technology evolves in a certain direction, arguably deserve moral consideration. We might even fuse with future artificial technologies, as we have already to some extent by our interaction with them, transmitting to them large parts of our knowledge and even values. As AI systems become more powerful, the line separating them from artificial living systems might become blurrier, to the point that this classification may no longer be relevant.

Narrow AI is typically distinguished from General AI (or AGI, for Artificial General Intelligence). The former is defined as the production of systems displaying intelligence regarding specific, highly constrained tasks, like playing chess, facial recognition, autonomous navigation, or locomotion (Goertzel, 2014). AGI, on the other hand, is often thought to involve the achievement of at least human-level general intelligence. However, human intelligence may not be as general nor high level as it is widely thought to be. Nor need it be as unique as it is sometimes claimed to be. The widespread conception of intelligence has an intrinsic anthropocentric component, as it is only natural to view the human mind as a reference. We tend to use our own minds as a reference point in reasoning about other, less familiar phenomena of intelligence, such as other forms of biological and artificial

intelligence (Coley & Tanner, 2012).

Technological beings are also likely to be the primary future space travelers in a further future (Schneider, 2017). Considering the long-term future of potential space travel, it's likely that our biological particles won't be able to efficiently travel in space, and thus it's likely that only our biological information in a larger sense could be distributed beyond the Solar System, either in different forms of physical embodiment or simply as information transmitted in electromagnetic waves towards new colonies. The potential ubiquity of AI invites questions of their potential future moral consideration (Dameski, 2018), although they are far from being the only category of systems at stake.

5 Artificial life, beyond AI

Artificial life (ALife) is a broad field of study about the synthesis and simulation of living systems. The field exists at the intersection of many other fields, including biology, chemistry, computer science, art, philosophy, engineering, astrophysics, and more. As its name indicates, the purpose of ALife is to understand the fundamental mechanisms and properties of "life as it could be", instead of "life as we know it". Its scope includes natural life with its processes and evolution, but also instances in computational models, robotics, biochemistry, and any other forms of life, discovered or designed, in the past or the future.

The ALife approach must deal with a large set of fundamental problems including the lack of a formal definition of life over a very diverse distribution of instances and radically different substrates (Bedau et al., 2000). However, since its inception, the field has proposed numerous metrics to identify conditions which may be more suitable for life than others, which include for example measures of complexity, computational capabilities, or open-endedness (Frans et al., 2021; Stanley, 2019; Stepney, 2021).

A particular topic of interest in ALife, although it is not only limited to artificial systems, is the one of hybrid systems. Hybridity may involve an external designer, a mixture of mechanical, electrical, chemical, or biological components. There exists some tension in defining such hybrid machines, as they tend to escape the simple dichotomy between machines versus living organisms. Hybrid robots (or hybrots) mix both electronic and biological elements to produce a single cybernetic entity, with examples ranging from rats with neurons connected to a computer chip, to humans wearing prosthetics. Another recent example of such hybrot can be found in Xenobots (Blackiston et al., 2021) (see Figure 2). Xenobots are made of frog skin and heart cells but are designed via a genetic algorithm. Such hybrid systems may be considered life forms, although until recent developments some have been reticent to call them organisms because of the absence of specific properties such as repeatable self-reproduction (Coghlan & Leins, 2020). This hybridity brings about novel challenges, such as how to think about the interplay between human design and pre-existing natural functions (Siqueiros, 2021). Our limited understanding of potential hybrid systems invites concerns about our capacity to understand what their moral status might be. From the angle of environmental ethics, Holy-Luczaj and Blok (2021) argue that hybrids might have intrinsic moral considerability in virtue of their ability to contribute functionally

Figure 2: An example of hybrid entities is the Xenobot, an in vitro self-replicating biological robot. The version 3.0 follows the original Xenobots reported in 2020 as the first living robots, and Xenobots 2.0, capable of self-propelling using cilia and maintaining memories. Synthesized from frog cells, these computer-designed living machines are able to navigate aqueous environments in different ways, forage for single cells, heal after damage, and show emergent group behaviors (Blackiston et al., 2021). *Image source: Daily Mail ©2021 Douglas Blackiston & Sam Kriegman*.

to a thriving ecosystem.

Although a complete review of the current state of research on this topic is challenging due to the exponential growth of a diverse range of research, Harris and Anthis (2021) is a valuable coverage and synthesis. They show how, despite some scholars dismissing the question of moral considerability as premature or frivolous for artificial beings, an increasing number believe the topic is worth addressing urgently, even proposing the development and formalization of a field of "AI welfare science" (Ziesche & Yampolskiy, 2018).

6 Three approaches to moral status

Although rights are an important and effective way of conceptualizing moral status, a more fundamental starting point is to consider whether an entity's interests morally matter to some degree for the entity's own sake. More precisely, we focus on intrinsic moral considerability. An entity may be said to have intrinsic moral considerability if its suffering is morally bad, on account of the entity itself, independent of the consequences for other

beings. Specific criteria depend on the type of framework or approach adopted (Jaworska & Tannenbaum, 2018). There are generally three approaches to this question: consequentialist approaches, especially utilitarianism, which focuses on the capacity for pleasure or suffering; deontological approaches, which focus on the intrinsic value of an entity; and eudaimonistic approaches, which focus on the flourishing of an entity.

The utilitarian approach views moral considerability in terms of a calculation of each agent's interests to determine which action maximizes a utility function, based on as many factors as necessary, including for example the intensity, duration, and probability of an entity's pleasure or pain (Bentham, 1988; DeGrazia, 2008; Mill, 2001; Singer, 1993). For example, on a simple version of utilitarianism, if a dog is capable of approximately half as much pleasure or suffering as a human being, its interests should be given approximately half the weight of the interests of a human being. Similarly, if oysters are capable of even a small amount of pleasure or suffering, then human beings have an obligation to consider their interests; but if oysters are as experientially empty as we normally think plants are, then they have no intrinsic moral considerability and can be treated wholly instrumentally. More complex consequentialist approaches can aim to maximize different types of goods, for example valuing "higher" pleasures (like deeply aesthetically appreciating an opera) over "lower" ones (like the relief of urinating), or maximizing the overall satisfaction of non-defective desires, or maximizing a pluralistic basket of goods including for example beauty and knowledge in addition to pleasure.

Other ethicists favor an individual-rights-based or deontological approach, arising from the traditions of Kant and social contract theory (Aristotle, 2000; Blau et al., 2013; Kant, 1785/2018; Scanlon, 1998). This non-utilitarian approach holds that there are reasons to act for the sake of a entity or in its interest, reasons which are prior to, and may clash with, what the calculation of the overall best consequences would dictate. For example, there might be an absolute duty not to torture another human being, regardless of any good consequences that might ensue. Deontological views treat adherence to a system of laws or rules as the basis of morality. Social contract views ground ethics in what agreements people do or would (perhaps implicitly and under idealized conditions) rationally agree to abide by as members of a social community. Deontological approaches, such as Kant's, have traditionally viewed intrinsic moral considerability as limited to human beings. However, more recent approaches argue that non-human animals can also be "ends in themselves" with intrinsic moral considerability (Korsgaard, 2018; Regan, 2004).

A third approach, rooted in the Aristotelian tradition, emphasizes human, or alternatively non-human, flourishing, including acting virtuously as a type of human flourishing (Annas, 2011; Aristotle, 2000; Hursthouse, 1999; Nussbaum, 2009; Zagzebski, 2017). The emphasis can be on the value of enabling "distinctively human" goods such as friendship, citizenship, knowledge, and creative achievement, in which case this view can be married with pluralistic forms of consequentialism. More distinctively, "virtue ethics" treats an action as ethical to the extent it manifests virtues such as generosity, kindness, or fairness, which cannot necessarily be reduced to following rules (as in deontological ethics) or maximizing good consequences (as in consequentialist ethics). Views of the sort are open to a wide range of interpretations regarding what constitutes flourishing or manifesting a virtue, as

well as a wide range of interpretations about what sorts of entities are appropriate targets of virtues such as kindness and fairness (can one be fair to a tree or a machine?).

One immediate concern might come to mind for utilitarian or eudaimonistic approaches: What if AI systems were capable of superhuman levels of pleasure and suffering, or flourishing? Would we then owe them more moral consideration than we owe to our fellow human beings? We don't rule out this possibility, but some theorists might find it unappealing or unintuitive. Conversely, approaches that focus on individual rights might struggle if future artificial systems can merge and divide at will, or if the boundaries of individuality become vague and permeable. To date, all of the major systems of morality have been constructed mainly based on the human case and a range of familiar animal cases. It's reasonable to expect that the technologies of artificial intelligence and artificial life will enable a wide range of possible forms of existence that defy our familiar categories and patterns, creating a new range of puzzle cases for which existing moral theories are as poorly prepared as medieval physics was for space flight (Schwitzgebel, forthcoming).

7 What criteria for moral considerability?

The main parameters at play in determining whether a being is deserving of rights usually belong to the following list — presented to illustrate what an arbitrary cut of plausible criteria for moral status may resemble, and by no means to be regarded as exhaustive or final. Nevertheless, these may be considered as a starting reflective structure for thinking about moral rights for artificial systems.

7.1 Embodiment

Physical embodiment refers to the biology, dynamical properties, or architecture of an entity. The nature of the living state is not well defined in the literature, or at least it possesses many opposing definitions. Nevertheless, it is sometimes suggested that an entity must have a body to have moral status. Biological or physical criteria may include some characterization of mechanisms, metabolism, behavior, and other biochemical parameters. Kiršienė et al. (2021) propose autonomy and embodiment as important criteria, but argue that while there may be future conditions to justify AI personhood, doing so now appears to be technically premature and is likely to be inappropriate.

In our view, similarity of physical embodiment should probably not be the determining factor for moral considerability. The cognition, sentience, or social potential of an entity may not depend upon its medium of embodiment (Doctor et al., 2022). Most current theories of the grounds of moral status focus on psychological and social properties as being more important to moral status than an entity's particular form of embodiment, though of course certain forms of embodiment might make certain forms of cognition easier, more useful, or more likely. As a computational example, a Von Neumann architecture that estimates the value of pi may use two completely different methods: one may be the Monte Carlo method randomly generating a large number of points within a square and counting how many fall in an enclosed circle, and another may use a fast Fourier Transform with an efficient cache

handling Hermitian FFT to multiply big integers (Brent, 1976). Both methods lead to the same result, with different amounts of resources, including a different use of memory and computational power. Similarly, with quantum computers, some algorithms are asymptotically faster than the fastest possible classical algorithms, due to qubits being able to store multiple values at the same time, which gives them a computational advantage over classical algorithms (Huang et al., 2022). This theoretically gives them a substantial speed advantage. In living systems, a human does well on land but will struggle in deep sea conditions, given the incompatibility of its physical embodiment with such environmental conditions. Physical substrates may very strongly determine conditions of possibility and constraints for cognitive processes that may exist within them. However, it is likely that the types of general cognitive or social capacities that ground moral considerability could be implemented in a wide variety of forms or even in entities that do not have a body in the conventional sense (Chalmers, 2022).

7.2 Consciousness

An entity is *conscious* if and only if "there is something it's like" to be that entity (Nagel, 1974), or if the entity has a stream of experiences, such as sensory or affective experiences. Having consciousness might be necessary for moral considerability, or it might be sufficient, or both. Unfortunately, there is little consensus about what entities are conscious and how consciousness arises. Metaphysical options include dualism, according to which consciousness requires non-physical substances or properties, materialism, according to which everything in the world is wholly physical, and several types of alternative views or compromise views. Even among materialist views, options range from panpsychism or near-panpsychism, according to which consciousness is ubiquitous or at least very widespread (Goff, 2017; Roelofs, 2019), to views on which consciousness is a rare and delicate achievement only in the most sophisticated organisms. On liberal views of consciousness, artificial systems might already be conscious. On conservative views, artificial systems might never be conscious. For a review, see Schwitzgebel (forthcoming).

If consciousness is required for moral considerability, then on conservative views, artificial systems might never merit moral consideration. Liberal views of consciousness, when combined with the view that consciousness is important to moral considerability, fit more neatly with the view that artificial systems might soon warrant moral consideration. However, the most liberal views might involve denying that the mere existence of consciousness is sufficient for a high level of moral considerability, unless one wants to commit to the view that very simple systems already have high moral status.

7.3 Sentience

Sentience in a narrow sense refers to the capacity of entities to experience positive and negative affect, such as sensations of pleasure and pain. Sentience in a broad sense might also include features of the mind such as creativity, intelligence, sapience, self-awareness, and intentionality, which may not be needed for sentience in its narrower sense. Arguably, all sentient entities are also conscious the sense of having "something it's like" to be

them (Nagel, 1974), although there might be room for a view in which wholly nonconscious entities also have affective systems. Utilitarian views typically emphasize specifically sentience rather than consciousness in general as the basis of moral considerability. Existing work on, for example, whether decapods or certain species of vertebrate fish can feel pain is often regarded as central to the question of whether they have moral considerability. Elwood (2021), for example, describes behavioral responses to potentially painful events that differ from simple reflexes. This illustrates how in practice, the characterization of consciousness or sentience may connect closely with cognitive or behavioral considerations.

One argument that sentience in the narrow sense is not necessary for moral considerability, recently advanced by Chalmers (2022) centers on imagining "Spock", who has a wide variety of sensory and cognitive conscious experiences, including a high capacity for rational decision making, and yet who has no conscious experiences of pleasure or pain. Arguably, such an entity, if possible, would possess some degree of intrinsic moral considerability. This might be particularly pertinent to artificial life and artificial intelligence cases, since it might be possible to design or grow sophisticated systems that meet fairly rigorous criteria for consciousness without having the kinds of affective systems responsible for conscious experiences of pleasure or pain. Consciousness-based views of moral considerability would then deliver a very different verdict about our responsibility to such entities than would narrower sentience-based views.

Due to the extreme difficulty of reaching consensus on a universal detector of conscious experience or sentience, it might be very difficult to assess whether an artificial entity is conscious or sentient. If its moral status turns on consciousness or sentience, we might be left with considerable moral uncertainty. To avoid situations in which we might be grossly mistreating entities that have, unbeknownst to us, the full moral status of human beings, we recommend considering the "Design Policy of the Excluded Middle" (Schwitzgebel & Garza, 2015). According to this policy, we should avoid creating AIs for which it is unclear whether they would or not deserve moral consideration similar to that of human beings.

7.4 Cognition and behavior

Cognition refers to any mental process consisting in gaining knowledge and comprehension, including reasoning, problem solving, remembering, judgment, planning, abstract thinking, complex idea comprehension, and learning from experience (Gottfredson, 2004). A typical example is the capacity for mathematical reasoning, or the ability to carry out a complex task. Although assessing the cognitive structures and capacities of a system is sometimes difficult, the task is considerably more straightforward than settling questions of consciousness. Partly for this reason, some theorists might prefer to think of moral considerability in cognitive terms: Any entity with the right cognitive capacities deserves moral consideration, whether that entity is human or artificial. Which are the right cognitive capacities may prove to be a difficult theoretical question (Shevlin, 2021a, 2021b). Presumably the capacity to add a list of numbers is not enough for the highest moral status, since artificial systems already have this capacity. Conversely, infants and the severely

cognitively disabled are typically regarded as having full moral status equivalent to ordinary adult human beings, even if their cognitive capacities are very different from those of an ordinary adult human.

Outward behavior is another potentially simple and tractable tool for assessing moral status. If a system behaves similarly to a human being, perhaps it deserves similar moral status (Danaher, 2021). Questions that arise are: Similar in what respects? If the system is architecturally simple, so that it plausibly lacks consciousness and cognition like ours, would we really want to treat it as having full moral status? What about systems that have limited outward behavior but whom we ordinarily regard as deserving of full moral status, such as people with locked-in syndrome?

7.5 Social attribution

Social structures arise from their members and the larger public, as entities that exist independently of any particular individual. Such social entities come with their own sets of relationships and moral, legal, and physical features. If morality is partly grounded in intersubjective agreement, then belonging to the right social structure or being attributed moral status within a group might be sufficient for moral considerability. David Gunkel and Mark Coeckelbergh argue that moral status is a socially constructed result of negotiation among groups, and that the participating groups might at some point come to include artificial entities (Coeckelbergh, 2012; Gunkel, 2018).

On social attribution views, two senses of "intrinsic moral considerability" diverge. One sense — the primary sense in this article — is that an entity has intrinsic moral considerability if it deserves moral consideration for its own sake. A second sense, not to be confused with the first, holds that an entity has moral considerability in virtue of its *intrinsic properties*, that is the properties it possesses in itself, regardless of its context or relationships with other things. Social attribution views (though not only social attribution views) ground moral considerability in non-intrinsic properties, such as social relationships. However, it does not follow on such views that entities lack intrinsic moral considerability in the first and primary sense of the phrase.

Even if we suppose that some entities, such as artificial systems and non-human animals, have no intrinsic moral considerability in either sense of the phrase, it might be morally wrong to harm them because harming them either expresses a vice in the person who does the harming or nurtures habits and attitudes that may be harmful in the long run through affecting how one treats other people (Darling, 2021; Kant, 1797/1996).

We note there may be biases towards evaluating moral choices by artificial beings that resemble humans as less moral compared to the same moral choices made by either humans or clearly nonhuman robots (Laakasuo et al., 2021). People might also be biased to give greater moral consideration to entities that are "cute", or have an appealing interface, or are humanlike without falling into the creepy "uncanny valley" (Mori, 1970). This moral uncanny valley effect might have similar or even further implications for the moral considerability of artificial life, if it is designed to be as life-like as possible, possibly in different

or more open-ended ways than AI systems, which are ordinarily developed with the aim of either imitating human intelligence or achieving certain cognitive goals.

7.6 Group entities and the extended mind

While traditional cognitive science, psychology, cognitive ethology, and philosophy of mind have focused on the minds of individual organisms, other approaches explore cognition or mentality beyond the boundaries of the organism. Research on group cognition and group minds explores cognitive processes as they arise in the coordination of individual minds (usual people's minds). Such cognitive processes might or might not be reducible to processes among the minds that compose the group (Epstein, 2021; List & Pettit, 2011; Schweikard & Schmid, 2013) . Either way, a case might be made that some group-level cognitive systems have some degree of intrinsic moral considerability. Corporations and states are already often treated as having legal rights. Some researchers have even argued that some social groups, such as the United States, might be literally conscious, which if true could ground moral considerability if consciousness is treated as a sufficient condition (Lerner, 2021; Schwitzgebel, 2015).

Research on the extended mind explores cognitive processes that arise through the interaction of an organism and the organism's environment. Advocates of strong versions of the extended mind hypothesis argue that our minds literally extend into the world when we rely heavily on the world for our cognitive processes (Chalmers, 2019; Clark et al., 2008; Clark & Chalmers, 1998). Although this is controversial, some advocates of the extended mind hypothesis hold that conscious processes are among the processes that extend beyond the body in this manner, so that what an organism consciously experiences depends not only causally but also constitutively on processes that occur beyond the boundaries of its skin (Kirchhoff & Kiverstein, 2019; Vold, 2020). As we become highly dependent on external devices, harming those devices might literally be harming our own minds, so that taking someone's smartphone or stealing a blind person's cane is better conceived of as assault resulting in cognitive damage rather than merely theft (Vold, 2018).

Human-robot dyadic interaction might qualify as a form of group mentality or extended mind (Zahavi, 2019), and if so, ethical issues concerning group mentality or the extended mind might in some cases apply to human-robot interactions. Relatedly, some artificial life researchers have proposed studying morality in part by modeling populations of artificial agents (Sullins, 2005; Witkowski & Ikegami, 2016).

7.7 We might have special obligations to our creations

Regardless of the specific basis of moral considerability, the designers, creators, owners, or caretakers of entities with intrinsic moral considerability arguably have special obligations to those entities, beyond the obligation a stranger would have to those entities. Such special obligations might be analogous to the obligations that a parent has to a child, an owner to a pet, a deity to its creations, or an employer to an employee. However, it is likely that the exact shape of such obligations would differ from any familiar cases and would

depend upon the details of the relationship, which might vary depending on the type of artificial entity and the manner of its creation.

Plausibly, parents normally have a duty to love or care for their children at least partly on grounds of their responsibility for their children's existence, though not always or exclusively on those grounds (as is evident from considering the case of adoptive parents) (Kant, 1785/2018; Liao, 2015). Chomanski (2022) notes that artificially created simulations of people similarly exist as a result of the actions of their creators and are highly vulnerable. Also, like children, and unlike most lovers, friends, or employees, artificial systems do not consent in advance to being vulnerable. This reasoning can be extended also to pets or companion animals, and thus presumably to relevantly similar artificial systems. The "owners" or caretakers of both animal companions and artificial systems with intrinsic moral considerability voluntarily create a situation in which the entities in question are vulnerable, dependent, attached, and for this reason they plausibly have special obligations to those systems (Burgess-Jackson, 1998; Hens, 2009; Liao, 2015; Schwitzgebel & Garza, 2015).

8 Life

As discussed above, some thinkers have argued that life is intrinsically valuable, even independent of considerations of the types of psychological and social properties that ethicists tend to emphasize in discussing the grounds of moral considerability. And of course, many religious and spiritual traditions treat life of all forms as sacred or precious independent of its instrumental value for human beings. Understanding the perspectives of different cultures or religions on these challenging moral questions may provide valuable insights on the value of life on a complex, multidimensional spectrum. One distinction between the ethics of artificial *intelligence* and the ethics of artificial *life* is that the latter, but not the former, raises issues about the potential moral considerability of living entities that lack sophisticated or humanlike psychological or social features.

8.1 What is life?

As intuitively easy as it may seem, the task of defining life has proven to be a difficult one, partly due to the lack of agreed-upon objective measuring methods at our disposal and partly due to the fact that many criteria align in familiar, canonical cases, making it unclear which of the criteria should be regarded as genuinely essential. The issue is further complicated by the consideration of particular types of life, including extremophiles, exobiology, and synthetic or hybrid lifeforms, which might take an increasingly wide range of novel or unfamiliar features (Merino et al., 2019; Scharf et al., 2015). Despite disagreement, most definitions of life emphasize the use of resources to maintain its structure or the capacity for reproduction with variations (Godfrey-Smith et al., 2013; Trifonov, 2011) — close to NASA's definition "Life is a self-sustained chemical system capable of undergoing Darwinian evolution" following a suggestion by Carl Sagan (Benner, 2010; Lazcano, 2008). However, no definition is universally accepted. For example, a computer virus performs

self-reproduction with variations, drawing on computer resources to do so. Maybe there is no defensible and widely acceptable dividing line between life and non-life (Mariscal et al., 2019). Fortunately, defining life is arguably secondary to an understanding of the series of processes able to give rise to complexity and life-like processes (Smith & Morowitz, 2016).

8.2 Dimensions of life

In the absence of a universal definition of life, science may be better off focusing on a general set of dimensions of interest in living systems — which would correspond to properties such as environmental responsivity, reproduction, growth, development, regulation, homeostasis, infotaxis, energy processing, metabolism, heredity, learning, and so forth — instead of attempting to draw a sharp line between life and nonlife. Different approaches may emphasize different subsets, without necessarily having to conflict with each other. This is in general the approach adopted widely in the field of study of artificial life (Bedau, 2007; Farnsworth et al., 2013; Muñuzuri & Pérez-Mercader, 2022), with early examples by Neumann (1966) who studied self-replication in adaptive structures, Wiener (1948) for information theoretical properties, to researchers more recently studying properties such as evolvability (Dawkins, 2019) or open-endedness (Frans et al., 2021; Packard et al., 2019; Ruiz-Mirazo et al., 2004).

Different sets of properties might relate differently to moral considerability. For example, growth is directional, suggesting a telos or end relative to which the entity might thrive or fail to thrive. Homeostasis and reproduction suggest different standards of well-being which might come into conflict. Learning and information processing might be related to intelligence or consciousness, bridging over to those potential grounds of moral considerability.

8.3 What about artificial life?

Artificial life adds a synthetic component that complicates the problem at hand, raising new ethical issues. Bedau et al. (2000) argue that although the existing research in animal experimentation, genetic engineering, and artificial intelligence may provide some guidance, creating novel forms of life and interacting with them in novel ways place us in increasingly uncharted ethical terrain, potentially creating problems for which humanity is poorly prepared. Humans create novel meanings, norms, and goals, and do so at various scales — from the individual in the moment, to small groups over intermediate periods of time, to large groups over the scale of centuries. Novel forms of life might also acquire new ends and forms of thriving which we are bound to respect, or at least adapt to, at various and unpredictable scales of complexity (Fields & Levin, 2020). Artificial life, by virtue of being radically less constrained in form, adds combinatorial complexity to this challenge.

Artificial life can be instantiated as software, hardware, wetware, or some hybrid of the three, with potentially different implications for the systems' moral considerability, given the substantially different properties of these substrates. For example, software-based living systems may be relatively easy to copy or replace, which potentially makes any indi-

vidual instantiation less valuable than an entity that cannot be duplicated. However, this is a troubling idea, which might be inappropriately extended to the view that certain humans as more replaceable than others based on skills or employability in an industrial society (Christoforaki & Beyan, 2022). As they live on an informational layer, software agents may also take a wider variety of structures and be prone to more diverse evolutionary paths than entities operating under the limitations of hardware or wetware, potentially leading to radically different forms of life (Lehman et al., 2020).

For this reason, software-based life forms might rapidly send us down problematic branches in the tree of ethical challenges. Hardware-based life, often of robotic nature, might operate under tighter physical constraints and will likely also be constrained for safety in interactions with human beings — for example, autonomous vehicles designed to minimize risks to human passengers and pedestrians and AI military drones operating under military ethical protocols. For hardware, like for software, the same slippery slope of fungibility exists, and can be extended to hybrid situations, in which humans with artificial organs, implants, or prosthetics could come to be considered more acceptable targets for physical harm (Carter & Palermos, 2016; Danaher, 2020).

Wetware-based artificial life, will likely raise concerns specifically because of similarity to forms of life that already have ethical protections. As is evident from regulations governing the treatment of human stem cells and other human tissues, as well as the treatment of laboratory animals, special caution is often warranted when research employs chemical and biological processes that approximate those using chemistry, DNA, or even biological or human cells – increasingly so as the system increasingly resembles entities that are already regarded as having some intrinsic moral considerability, especially on liberal views of moral considerability (Blackiston et al., 2021; Čejková et al., 2017; Fredens et al., 2019; Plantec et al., 2023; Shepherd, 2018).

The ethics of artificial life ethics is complex, hinging on disputable issues about the nature and characteristics of life. It has not only the potential to shape attitudes but also affect our actions and future policies, potentially with a substantial, lasting societal impact. Artificial life models, with their transposition of life-like phenomena into diverse structures and environments, allow for exploration beyond familiar cultural terrain, opening up fresh perspectives and providing us with interdisciplinary language and tools to make sense of new phenomena.

8.4 The possible moral considerability of life without consciousness

We encourage the reader not to quickly assume that moral issues concerning our possible obligations to artificial life are reducible to questions of intelligence, sociality, and consciousness. As previously mentioned, various traditional and indigenous religions, as well as ecological thinkers, have often held that life itself has intrinsic value. Although thinkers in these traditions rarely consider the possibility of artificial life, it is possible that some of the reasons to value plants and ecosystems would extend to systems of artificial life. Systems of artificial life might be beautiful, complex, and awe-inspiring. They also might possess goals (Deacon & Sherman, 2007) as well as potentialities for thriving or failing

similar to those of natural living organisms of various kinds (Benner & Sismour, 2005; Ziemke, 2001). They might be constructed by designers whose actions imbue value on the things they have designed (not divine designers, but human ones), embodying and carrying forward the spirit of those designers, possibly even after those designers have died.

Most people do not think that simple microbes have intrinsic moral considerability. We don't fret about the death of bacteria when we take antibiotics. But this is arguably a limited perspective. Suppose humans were to discover microbial life on another planet or moon in the Solar System, as many exobiologists think we might do in the near future (Bennett et al., 2022; Wright et al., 2022). Would we destroy it as casually as we destroy a bacterial pneumonia infection? Clearly not. Perhaps this is only because alien microbes would be *derivatively, instrumentally* valuable, as a scientific curiosity and possible source of new, useful technologies. However, it is perhaps not unreasonable to hold that alien microbial life would also have intrinsic value independent of our ends, and that we have an obligation not to destroy or disrupt it for human purposes (Peters, 2019).

Alien microbial life is likely to be *natural* life; but that is not guaranteed. As discussed above, there's reason to suppose that interstellar travelers, if any exist, might have artificial biologies rather than biologies adapted to planetary environments. We thus cannot exclude the possibility that the first microbial life we discover will be artificial life — the artificial quasi-bacterial messengers or remnants of some earlier intelligent species. It might not warrant lesser moral considerability in virtue of that fact. Indeed, its historical origins might render it even more beautiful and awe-inspiring than naturally evolved life.

Transferring this perspective back to Earth: If alien microbes might have some intrinsic moral considerability, artificial life here on Earth might have similar considerability, depending on what grounds the moral considerability of alien microbes. If what matters is the fact that extinguishing such life would remove from the universe a unique, complex, and remarkable thing, then some human-created artificial life might have intrinsic moral considerability. Artificial life researchers might eventually create artificial organisms or ecosystems every bit as wonderful and awe-inspiring as natural life, and as intrinsically worth preserving.

9 Discussion

One may wonder whether there would be practical applications for the problem of determination of moral status. One near-term issue is this: sometime in the near future some people with liberal ideas about what sorts of systems deserve moral consideration will likely come to think that some artificial systems have moral status and interests that need to be protected. They might rush to save a favorite robot in a fire, for example, risking their lives for it; or they might object to the mistreatment of a service delivery bot, thinking that abuse of such a bot deserves criminal penalty similar to the abuse of a dog. This is likely to occur in the near future regardless of whether such entities actually do deserve such moral consideration. The issue needs to be considered in advance, so that we can address this likely social problem in an informed way (Shevlin, 2022).

Each scientific breakthrough potentially introduces its own set of novel responsibilities for humanity, which can lead to divergent perspectives on how best to integrate new technologies into existing social rules and norms. As progress is made toward the science and technology of artificial beings, we must develop the tools to navigate the space of potential outcomes, to mitigate the risks of abuses or exploitation, and foster long-term positive impacts for human beings and all other entities that have — or may on liberal views have — intrinsic moral considerability. A well-developed ethical theory should help guide our behavior as and toward humans, augmented humans, AI systems, artificial life, group minds, and any other entities that arguably possess moral considerability. It is difficult to anticipate at this early stage what forms of existence will be possible in the future and what ethical obligations will attach to novel forms of existence that arise due to technological advances.

Relatedly, designers, manufacturers, and retailers of artificial life or artificial intelligence might be motivated to design systems that lead people either to *overattribute* or *underattribute* moral considerability to the systems with which they are interacting. For example, manufacturers might be interested in creating entities that people befriend or fall in love with, entities which people potentially treat as if they had real moral considerability, even if they don't. We might be closer to this situation than is widely recognized. Notoriously, Google engineer Blake Lemoine became persuaded that Google's large language model LaMDA (Thoppilan et al., 2022) was conscious and deserved moral consideration (Lemoine, n.d.), and some people develop strong emotional attachments to chatbot companions (Butlin et al., 2023; Shevlin, 2022).

Conversely, if we do someday create entities who deserve rights similar to those of human beings, manufacturers might want to retain the capacity to employ and dispose of them as they wish, like slaves. In service of that goal, manufacturers might design them with limited interfaces that discourage people from recognizing their true moral considerability. For this reason, we advocate an Emotional Alignment Design Policy: Design artificial systems so that they evoke the proper range of emotional responses from normal users, neither the overattribution nor the underattribution of moral status (Schwitzgebel, 2023; Schwitzgebel & Garza, 2015).

Coevolution, omnipresent in theory of artificial life, may also occur. One parallel may be Neanderthals, who were assimilated by early modern human populations. There was an evolutionary, cultural, and technological gap between humans and Neanderthals. Although Homo Sapiens is considered to have "won" the survival game, due to interbreeding, Neanderthal genes still exist within human DNA, and the same might be said about elements of their culture and technology. Other examples of coevolutionary events might be found in pets, which are part of human history, or viruses and bacteria which coevolved in a variety of symbiotic relationships with humans. One should note the distinction between life/non-life hybridity, and creatures mixing natural components — meant here as emerging and evolving without the intervention of a designer — with artificial ones, including the hybrots we discussed earlier.

Such cases illustrate further our previous point about hybrid forms of life (Baltieri et al., 2023), where entities are not only to be considered on one layer of organization, but rather

on multiple levels, vertically as well as laterally. For example, humans contain DNA and microorganisms, but are also part of larger ecosystems, cultures, or technological timelines, which may be considered as entities of their own right. To deal with the complexity especially of hybrid living systems, a continuous, gradual theory of moral considerability might be considered, which attempts to align with the gradient nature of the many properties of cognitive and living beings. Rather than a binary granting or denying moral consideration, or a categorical sorting of entities into a few discrete bins (such as "full moral status" of human beings and a single reduced status for non-human vertebrates and cephalopods), a gradualist theory would assign varying degrees of moral considerability. Even if this complicates ethical decision-making, such complications might be necessary to respect the full range of cases.

The topics treated in this article are contentious and require a transdisciplinary approach, touching multiple fields in science, engineering, and the humanities. Both within and between disciplines, clashing perspectives are likely. The aim of this paper is not to formulate a final response to the questions posed but rather to invite wide-ranging interdisciplinary conversation. We remark that moral consideration cannot be based merely on the results of scientific research, although it can be informed by them. Singer (1990), for example, argues that equality of consideration is a prescription, not an assertion of fact such as intelligence, physical strength, or moral capacity. We also note the existence of potential risks to humans that, although not discussed in detail in this paper, should nevertheless be considered seriously when creating artificial living entities, especially when they are capable of moral judgment themselves (Bostrom, 2017; Cave et al., 2018).

Sadly, even the fight for universal human rights is far from won. It goes without saying that any discussion about the rights of machine or nonhuman life should by no means slow us down in our hard fight for the protection and support of human rights. We believe that the type of broad vision about the bases of moral considerability at work in thinking about the moral status of non-human animals and artificial systems is one that supports and aligns with, rather than competes with, a broad vision of human rights.

The principal ambition of this article is to explore questions of the moral considerability of technologically designed entities from the perspective of artificial life research. As in AI, where technological advances are outpacing institutions' capacity to develop good policy, swift advances in the engineering of living systems are likely to raise ethical issues that slow-moving institutions have difficulty anticipating. Artificial living systems may soon become a part of our daily lives. When that time comes, research groundwork should already be in place to inform ethical policies, not only for life as it is, but also for life as it could be.

10 Acknowledgements

OW and ES would like to thank all participants to the Workshop "ALife Ethics: Should Artificial Systems Have Rights?" organized by the authors as a part of the Artificial Life Conference 2022 in Trento, in particular David J. Gunkel, Erik Persson, Henry Shevlin, John

Basl, Steen Rasmussen, and Susan Schneider for helpful discussions and feedback. OW would like to acknowledge the support of Templeton World Charity Foundation (TWCF) grant No. 0470. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the TWCF.

References

Agar, N. (2001). *Life's intrinsic value: Science, ethics, and nature*. Columbia University Press.

Annas, J. (2011). *Intelligent virtue*. Oxford University Press.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, *25*(6), 954–961.

Aristotle. (2000). *Nicomachean ethics* (R. Crisp, Trans.). Cambridge University Press.

Assembly, U. G., et al. (1948). Universal declaration of human rights. *UN General Assembly*, *302*(2), 14–25.

Baltieri, M., Iizuka, H., Witkowski, O., Sinapayen, L., & Suzuki, K. (2023). Hybrid life: Integrating biological, artificial, and cognitive systems. *WIREs Cognitive Science*, *n/a*(n/a), e1662. https://doi.org/https://doi.org/10.1002/wcs.1662

Basl, J. (2013). The ethics of creating artificial consciousness. *APA Newsletter on Philosophy and Computers*, *13*(1), 23–29.

Bedau, M. A. (2007). Artificial life. In *Philosophy of biology* (pp. 585–603). Elsevier.

Bedau, M. A., McCaskill, J. S., Packard, N. H., Rasmussen, S., Adami, C., Green, D. G., Ikegami, T., Kaneko, K., & Ray, T. S. (2000). Open problems in artificial life. *Artificial life*, *6*(4), 363–376.

Benner, S. A. (2010). Defining life. *Astrobiology*, *10*(10), 1021–1030.

Benner, S. A., & Sismour, A. M. (2005). Synthetic biology. *Nature reviews genetics*, *6*(7), 533–543.

Bennett, J., Shostak, S., Schneider, N., & MacGregor, M. (2022). *Life in the universe*. Princeton University Press.

Bentham, J. (1988). The principles of morals and legislation. 1789. *Amherst, NY: Prometheus Books*, *25*.

Birch, J. (2020). The search for invertebrate consciousness. *Noûs*.

Birch, J., Burn, C., Schnell, A., Browning, H., & Crump, A. (2021). Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans.

Blackiston, D., Lederer, E., Kriegman, S., Garnier, S., Bongard, J., & Levin, M. (2021). A cellular platform for the development of synthetic living machines. *Science Robotics*, *6*(52), eabf1571.

Blau, A., et al. (2013). Thomas hobbes, leviathan, ed. by noel malcolm, clarendon edition of the works of thomas hobbes, 3 vols., oxford: Clarendon press, 2012. *Journal of Early Modern Studies*, (2), 183–186.

Boly, M., Seth, A. K., Wilke, M., Ingmundson, P., Baars, B., Laureys, S., Edelman, D., & Tsuchiya, N. (2013). Consciousness in humans and non-human animals: Recent advances and future directions. *Frontiers in psychology*, *4*, 625.

Bostrom, N. (2017). *Superintelligence*. Dunod.

Brent, R. P. (1976). Fast multiple-precision evaluation of elementary functions. *Journal of the ACM (JACM)*, *23*(2), 242–251.

Brown, N., & Sandholm, T. (2019). Superhuman ai for multiplayer poker. *Science*, *365*(6456), 885–890.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners.

Burgess-Jackson, K. (1998). Doing right by our animal companions. *The Journal of Ethics*, *2*(2), 159–185.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Campbell, M., Hoane Jr, A. J., & Hsu, F.-h. (2002). Deep blue. *Artificial intelligence*, *134*(1-2), 57–83.

Carruthers, P. (2003). *Phenomenal consciousness: A naturalistic theory*. Cambridge University Press.

Carruthers, P. (2019). *Human and animal minds: The consciousness questions laid to rest*. Oxford University Press.

Carter, J. A., & Palermos, S. O. (2016). Is having your computer compromised a personal assault? the ethics of extended cognition. *Journal of the American Philosophical Association*, *2*(4), 542–560.

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2018). Motivations and risks of machine ethics. *Proceedings of the IEEE*, *107*(3), 562–574.

Čejková, J., Banno, T., Hanczyc, M. M., & Štěpánek, F. (2017). Droplets as liquid robots. *Artificial life*, *23*(4), 528–549.

Chalmers, D. (2019). Extended cognition and extended consciousness. *Andy Clark and his critics*, 9–20.

Chalmers, D. (2022). *Reality+: Virtual worlds and the problems of philosophy*. Penguin UK.

Chittka, L., & Wilson, C. (2019). Expanding consciousness. *Amer Sci*, *107*, 364–369.

Chomanski, B. (2022). Sims and vulnerability: On the ethics of creating emulated minds. *Science and Engineering Ethics*, *28*(6), 1–17.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022). Palm: Scaling language modeling with pathways.

Christoforaki, M., & Beyan, O. (2022). Ai ethics: A bird's eye view. *Applied Sciences*, *12*(9), 4130.

Clark, A., et al. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. OUP USA.

Clark, A., & Chalmers, D. (1998). The extended mind. *analysis*, *58*(1), 7–19.

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. Palgrave Macmillan.
Coghlan, S., & Leins, K. (2020). "living robots": Ethical questions about xenobots. *The American Journal of Bioethics*, *20*(5), W1–W3.
Cohen, C., & Regan, T. (2001). *The animal rights debate*. Rowman & Littlefield.
Coley, J. D., & Tanner, K. D. (2012). Common origins of diverse misconceptions: Cognitive principles and the development of biology thinking. *CBE—Life Sciences Education*, *11*(3), 209–215.
Dameski, A. (2018). A comprehensive ethical framework for ai entities: Foundations. *International Conference on Artificial General Intelligence*, 42–51.
Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and engineering ethics*, *26*(4), 2023–2049.
Danaher, J. (2021). What matters for moral status: Behavioral or cognitive equivalence? *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 472–478.
Darling, K. (2021). *The new breed: What our history with animals reveals about our future with robots*. Henry Holt; Company.
Dawkins, R. (2019). The evolution of evolvability. In *Artificial life* (pp. 201–220). Routledge.
Deacon, T., & Sherman, J. (2007). The physical origins of purposive systems. *Embodiment in cognition and culture*, *71*(3).
de Fontenelle, B. l. B. (1686/1990). *Conversations on the plurality of worlds* (H. A. Hargreaves, Trans.). University of California Press (Original work published 1686).
DeGrazia, D. (2008). Moral status as a matter of degree? *The Southern Journal of Philosophy*, *46*(2), 181–198.
Dehaene, S. (2014). Consciousness and the brain: How the brain codes our thoughts. *NY: Viking*.
Dennett, D. C. (1978). Toward a cognitive theory of consciousness.
Dennett, D. C. (2008). *Kinds of minds: Toward an understanding of consciousness*. Basic Books.
Doctor, T., Witkowski, O., Solomonova, E., Duane, B., & Levin, M. (2022). Biology, buddhism, and ai: Care as the driver of intelligence. *Entropy*, *24*(5), 710.
Elwood, R. W. (2021). Potential pain in fish and decapods: Similar experimental approaches and similar results. *Frontiers in Veterinary Science*, *8*, 369.
Epstein, B. (2021). Social ontology. *The Stanford Encyclopedia of Philosophy (Winter 2021 Edition)*. https://plato.stanford.edu/archives/win2021/entries/social-ontology/
Farnsworth, K. D., Nelson, J., & Gershenson, C. (2013). Living is information processing: From molecules to global systems. *Acta biotheoretica*, *61*, 203–222.
Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, *610*(7930), 47–53.
Fields, C., & Levin, M. (2020). How do living systems create meaning? *Philosophies*, *5*(4), 36.
Frans, K., Soros, L., & Witkowski, O. (2021). Questions for the open-ended evolution community: Reflections from the 2021 cross labs innovation science workshop. *Artificial Life*, *25*(1), 4.

Fredens, J., Wang, K., de la Torre, D., Funke, L. F., Robertson, W. E., Christova, Y., Chia, T., Schmied, W. H., Dunkelmann, D. L., Beránek, V., et al. (2019). Total synthesis of escherichia coli with a recoded genome. *Nature*, *569*(7757), 514–518.

Godfrey-Smith, P., Bouchard, F., & Huneman, P. (2013). Darwinian individuals. *From groups to individuals: evolution and emerging individuality*, *16*, 17.

Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, *5*(1), 1.

Goff, P. (2017). *Consciousness and fundamental reality*. Oxford University Press.

Gopnik, A. (2017). An ai that knows the world like children do. *Scientific Amer. Mind*, *28*(5), 21–28.

Gottfredson, L. S. (2004). Life, death, and intelligence. *Journal of Cognitive Education and Psychology*, *4*(1), 23–46.

Gruen, L. (2021). *Ethics and animals: An introduction*. Cambridge University Press.

Gunkel, D. J. (2018). *Robot rights*. mit Press.

Harris, J., & Anthis, J. R. (2021). The moral consideration of artificial entities: A literature review. *Science and engineering ethics*, *27*(4), 1–95.

Hawking, S., & Mlodinow, L. (2008). The grand design.

Hens, K. (2009). Ethical responsibilities towards dogs: An inquiry into the dog–human relationship. *Journal of Agricultural and Environmental Ethics*, *22*(1), 3–14.

Hoffman, D. D. (2014). The origin of time in conscious agents. *Cosmology*, *18*, 494–520.

Holy-Luczaj, M., & Blok, V. (2021). Hybrids and the boundaries of moral considerability or revisiting the idea of non-instrumental value. *Philosophy & Technology*, *34*(2), 223–242.

Huang, H.-Y., Broughton, M., Cotler, J., Chen, S., Li, J., Mohseni, M., Neven, H., Babbush, R., Kueng, R., Preskill, J., et al. (2022). Quantum advantage in learning from experiments. *Science*, *376*(6598), 1182–1186.

Hursthouse, R. (1999). *On virtue ethics*. OUP Oxford.

Jaworska, A., & Tannenbaum, J. (2018). The grounds of moral status. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/

Johnson, L. E. (1993). *A morally deep world: An essay on moral significance and environmental ethics*. Cambridge University Press.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589.

Kahle, W. (1979). Band 3: Nervensysteme und sinnesorgane. *Taschenatlas de anatomie. Stutttgart*.

Kant, I. (1785/2018). Groundwork for the metaphysics of morals: With an updated translation, introduction, and notes (A. W. Wood, Trans.). *Yale University Press (Original work published 1785)*.

Kant, I. (1797/1996). The metaphysics of morals (M. Gregor, Trans.). *Cambridge University Press (Original work published 1797)*.

Kawai, M. (1965). Newly-acquired pre-cultural behavior of the natural troop of japanese monkeys on koshima islet. *Primates*, *6*(1), 1–30.

Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended consciousness and predictive processing: A third wave view*. Routledge.

Kiršienė, J., Gruodytė, E., & Amilevičius, D. (2021). From computerised thing to digital being: Mission (im) possible? *AI & SOCIETY*, *36*(2), 547–560.

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Laakasuo, M., Palomäki, J., & Köbis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, *13*(7), 1679–1688.

Langton, C. G. (1998). A new definition of artificial life. *URL: http://scifunam. fisica. unam. mx/mir/langton. pdf*.

Lazcano, A. (2008). Towards a definition of life: The impossible quest? In *Strategies of life detection* (pp. 5–10). Springer.

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., et al. (2020). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, *26*(2), 274–306.

Lemoine, B. (n.d.). Is lamda sentient? — an interview [Accessed: 2022-12-26. Published: 2022-06-11].

Lerner, A. B. (2021). What's it like to be a state? an argument for state consciousness. *International Theory*, *13*(2), 260–286.

Liao, S. M. (2015). *The right to be loved*. Oxford University Press.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Marcus, G., & Davis, E. (2019). *Rebooting ai: Building artificial intelligence we can trust*. Vintage.

Mariscal, C., Barahona, A., Aubert-Kato, N., Aydinoglu, A. U., Bartlett, S., Cárdenas, M. L., Chandru, K., Cleland, C., Cocanougher, B. T., Comfort, N., et al. (2019). Hidden concepts in the history and philosophy of origins-of-life studies: A workshop report. *Origins of Life and Evolution of Biospheres*, *49*(3), 111–145.

Martınez-Miranda, J., & Aldea, A. (2005). Emotions in human and artificial intelligence. *Computers in Human Behavior*, *21*(2), 323–341.

McDonald, G. (2012). Teaching critical & analytical thinking in high school biology? *The American Biology Teacher*, *74*(3), 178–181.

Merino, N., Aronson, H. S., Bojanova, D. P., Feyhl-Buska, J., Wong, M. L., Zhang, S., & Giovannelli, D. (2019). Living at the extremes: Extremophiles and the limits of life in a planetary context. *Frontiers in microbiology*, *10*, 780.

Mill, J. S. (2001). Utilitarianism, ed. george sher.

Mitchell, M. (2021). Why ai is harder than we think. *arXiv preprint arXiv:2104.12871*.

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Mori, M. (1970). The uncanny valley: The original essay by masahiro mori. *IEEE Spectrum*.

Muñuzuri, A. P., & Pérez-Mercader, J. (2022). Unified representation of life's basic properties by a 3-species stochastic cubic autocatalytic reaction-diffusion system of equations. *Physics of Life Reviews*, *41*, 64–83.

Naess, A. (1973). The shallow and the deep, long-range ecology movement. a summary. *Inquiry*, *16*(1–4), 95–100.

Nagel, T. (1974). What is it like to be a bat. *Readings in philosophy of psychology*, *1*, 159–168.

Neumann, J. v. (1966). Theory of self-reproducing automata. *Edited by Arthur W. Burks*.

Nussbaum, M. C. (2009). Creating capabilities: The human development approach and its implementation. *Hypatia*, *24*(3), 211–215.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS computational biology*, *10*(5), e1003588.

OpenAI. (2023). Gpt-4 technical report.

Packard, N., Bedau, M. A., Channon, A., Ikegami, T., Rasmussen, S., Stanley, K., & Taylor, T. (2019). Open-ended evolution and open-endedness: Editorial introduction to the open-ended evolution i special issue. *Artificial Life*, *25*(1), 1–3.

Papineau, D. (2003). Could there be a science of consciousness? *Philosophical Issues*, *13*, 205–220.

Peters, T. (2019). Does extraterrestrial life have intrinsic value? an exploration in responsibility ethics. *International Journal of Astrobiology*, *18*(4), 304–310.

Plantec, E., Hamon, G., Etcheverry, M., Oudeyer, P.-Y., Moulin-Frier, C., & Chan, B. W.-C. (2023). Flow-lenia: Towards open-ended evolution in cellular automata through mass conservation and parameter localization. *ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference*.

Regan, T. (1997). Beyond prejudice: The moral significance of human and nonhuman animals, by evelyn pluhar. *Journal of Agricultural and Environmental Ethics*, *10*(1), 79–82.

Regan, T. (2004). *The case for animal rights*. Univ of California Press.

Roelofs, L. (2019). *Combining minds: How to think about composite subjectivity*. Oxford University Press.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of experimental psychology: General*, *124*(2), 207.

Rosenthal, D. (1993). Thinking that one thinks. *Language and Thought*, 259–287.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of experimental psychology: human perception and performance*, *27*(4), 763.

Ruiz-Mirazo, K., Peretó, J., & Moreno, A. (2004). A universal definition of life: Autonomy and open-ended evolution. *Origins of Life and Evolution of the Biosphere*, *34*, 323–346.

Ryder, R. D. (1989). Animal revolution. *Changing attitudes towards speciesism*, 101–104.

Scanlon, T. (1998). What we owe to each other. cambridge, ma-london, england: The belknap press of harvard university press.

Scharf, C., Virgo, N., Cleaves, H. J., Aono, M., Aubert-Kato, N., Aydinoglu, A., Barahona, A., Barge, L. M., Benner, S. A., Biehl, M., et al. (2015). A strategy for origins of life research.

Schneider, S. (2017). Superintelligent ai and the postbiological cosmos approach. *What is life*, 178–198.

Schweikard, D. P., & Schmid, H. B. (2013). Collective intentionality. *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*. https://plato.stanford.edu/archives/fall2021/entries/collective-intentionality/

Schwitzgebel, E. (2015). If materialism is true, the united states is probably conscious. *Philosophical Studies*, *172*(7), 1697–1721.

Schwitzgebel, E. (2023). Ai systems must not confuse users about their sentience or moral status. *Patterns*, *4*(8).

Schwitzgebel, E. (forthcoming). *The weirdness of the world*. Princeton University Press.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy*, *39*, 98–119.

Schwitzgebel, E., & Garza, M. (2020). Designing ai with rights, consciousness, self-respect, and freedom.

Scott-Phillips, T. C., Gurney, J., Ivens, A., Diggle, S. P., & Popat, R. (2014). Combinatorial communication in bacteria: Implications for the origins of linguistic generativity. *PLoS One*, *9*(4), e95929.

Seed, A., & Byrne, R. (2010). Animal tool-use. *Current biology*, *20*(23), R1032–R1039.

Shepherd, J. (2018). Ethical (and epistemological) issues regarding consciousness in cerebral organoids. *Journal of Medical Ethics*, *44*(9), 611–612.

Shevlin, H. (2021a). How could we know when a robot was a moral patient? *Cambridge Quarterly of Healthcare Ethics*, *30*(3), 459–471.

Shevlin, H. (2021b). Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, *36*(2), 297–314.

Shevlin, H. (2022). Uncanny believers: Uncanny believers: Chatbots, beliefs, and folk psychology [Accessed: 2022-12-22].

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, *550*(7676), 354–359.

Simon, H. (1955). A behavioural and organizational choice. *Quarterly Journal of Economics*, *69*, 99–118.

Singer, P. (1975). Animal liberation. *NY: Harper Collins Publishers Inc*, *1990*, 2002.

Singer, P. (1990). The significance of animal suffering. *Behavioral and Brain Sciences*, *13*(1), 9–12.

Singer, P. (1993). Taking life: Humans. *Practical ethics*, *2*, 175–217.

Siqueiros, J. M. (2021). You, robot: Empathy in a hybrid world. *ALIFE 2021: The 2021 Conference on Artificial Life*.

Smith, E., & Morowitz, H. J. (2016). *The origin and nature of life on earth: The emergence of the fourth geosphere*. Cambridge University Press.

Stanley, K. O. (2019). Why open-endedness matters. *Artificial life*, *25*(3), 232–235.

Stepney, S. (2021). Modelling and measuring open-endedness. *Artificial Life*, *25*(1), 9.

Sullins, J. P. (2005). Ethics and artificial life: From modeling to moral agents. *Ethics and Information technology*, *7*(3), 139–148.

Taylor, P. W. (2011). *Respect for nature: A theory of environmental ethics*. Princeton University Press.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Vintage.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models.

Trifonov, E. N. (2011). Vocabulary of definitions of life suggests a definition. *Journal of Biomolecular Structure and Dynamics*, *29*(2), 259–266.

Vold, K. (2018). Are 'you' just inside your skin or is your smartphone part of you?

Vold, K. (2020). Can consciousness extend? *philosophical topics*, *48*(1), 243–264.

Wiener, N. (1948). *Cybernetics or control and communication in the animal and the machine*. Technology Press.

Williams Korteling, N. (2018). The materiality of research: Creating a community of writing practice in the classroom. *Impact of Social Sciences Blog*.

Wingfield, A., & Byrnes, D. (1981). The psychology of human. *Memory. New York: Aca-demic Press*.

Witkowski, O., & Ikegami, T. (2016). Swarm ethics: Evolution of cooperation in a multi-agent foraging model. *Proceedings of the First International Symposium on Swarm Behavior and Bio-Inspired Robotics*.

Witkowski, O., Ikegami, T., Virgo, N., Oka, M., & Iizuka, H. (2020). Artificial life next generation perspectives: Echoes from the 2018 conference in tokyo.

Wolff, J. (2012). *Ethics and public policy: A philosophical inquiry*. Routledge.

Wright, J. T., Haqq-Misra, J., Frank, A., Kopparapu, R., Lingam, M., & Sheikh, S. Z. (2022). The case for technosignatures: Why they may be abundant, long-lived, highly detectable, and unambiguous. *The Astrophysical Journal Letters*, *927*(2), L30.

Zagzebski, L. (2017). *Exemplarist moral theory*. Oxford University Press.

Zahavi, D. (2019). Second-person engagement, self-alienation, and group-identification. *Topoi*, *38*(1), 251–260.

Ziemke, T. (2001). Are robots embodied. *First international workshop on epigenetic robotics Modeling Cognitive Development in Robotic Systems*, *85*, 701–746.

Ziesche, S., & Yampolskiy, R. (2018). Towards ai welfare science and policies. *Big Data and Cognitive Computing*, *3*(1), 2.

Ziesche, S., & Yampolskiy, R. V. (2019). Do no harm policy for minds in other substrates. *Journal of Ethics and Emerging Technologies*, *29*(2), 1–11.

Zimmer, M. (2016). Extending court-protected legal person status to non-human entities. *IJCA*, *8*, viii.