The Ethics of Modifying Artificial Sentient Beings

Kamil Mamak[1]

Abstract. If artificial sentient beings were to exist, how could we permissibly treat them? A natural assumption holds that the power to create such beings carries with it the right to modify them at will, reshaping them to serve our purposes. This paper challenges that assumption. I argue that the capacity to create does not confer an open-ended license to alter. Once a sentient artificial being exists, it has a standing of its own, and the default is that it should be left as it is, even where its behavior disappoints us. This default admits only narrow exceptions. An intervention requires genuine justification, of which the protection of others from harm is the clearest, and even then it must be targeted at the source of the danger rather than reshaping the being wholesale. Consent may widen this space, though whether an artificial subject could genuinely consent remains an open and difficult question. Throughout, one limit holds: every permissible modification must preserve the psychological continuity in which the subject consists. A change that breaches it is no longer modification but elimination, the ending of one subject in favor of, at most, a successor. Becoming an individual sentient being may thus entail a claim against radical alteration imposed by others.

1. INTRODUCTION

Sentience in the context of AI ethics is discussed in several distinct but related frameworks. One strand of the debate concerns whether we should create artificial entities capable of experiencing pain or suffering at all. Some scholars argue for caution. For example, Metzinger has called for a moratorium on the creation of artificial systems with phenomenal consciousness [1]. Others, however, defend a more permissive stance. For instance, recent work by Mamak explores potential advantages of creating artificial beings capable of suffering under carefully regulated conditions [2].

A second major line of debate concerns the criteria for granting moral status and the role sentience plays in meeting those criteria. Positions vary considerably, ranging from relational approaches see, e.g. , [3], [4], [5], [6], [7], [8], [9] to Kantian-inspired views see, e.g. , [10]. Nonetheless, the dominant position in contemporary AI ethics, remains a properties-based account according to which sentience functions as a sufficient, or at least central, condition for moral status see, e.g. , [11], [12], [13], [14].

A third context concerns the normative consequences that follow if artificial entities were to possess moral standing. Recognition of moral considerability is one issue; determining how such beings ought to be treated is another. These questions are often grouped under the headings of robot rights [8], [10], see, e.g. , [15], [16], [17], [18], [19] and AI welfare [20], [21], but see [22]. Within this debate, scholars distinguish between negative obligations, such as refraining from causing pain, and positive duties, such as providing energy, maintenance, or other forms of support necessary for continued existence and well-being see, e.g. , [11], [23], [24], [25].

This paper contributes to this third strand of the debate. It focuses specifically on the ethical permissibility of intervening in the mental lives of sentient artificial beings, and more precisely on the moral implications of modifying, enhancing, or otherwise altering artificial sentience. I argue that altering artificial sentient beings may, to some extent, be ethically permissible. However, there are likely to be limits to such interventions, especially when the proposed modification would amount to the practical elimination of the existing subject.

If certain forms of intervention would effectively extinguish the individual who bears moral status, this imposes normative constraints on how we may treat such entities. The emergence of sentient artificial beings may therefore generate obligations not only to avoid harming them, but also to respect their psychological continuity and individuality. In other words, becoming an individual sentient being may entail a claim against radical alteration imposed by others.

The paper is structured as follows. After these introductory remarks, I discuss the limits of intervention in the self of human beings. The paper then turns to the limits on modifying artificial sentient beings specifically. It closes with conclusions.

Before proceeding, I want to address the question of sentience and the reasons for discussing it. This topic is among the most polarizing issues in contemporary media coverage of the development of artificial intelligence. Some commentators hold that artificial entities are not, and

[1] Corresponding author: PhD, Jagiellonian University, Faculty of Law and Administration, Department of Criminal Law, Poland. Email: kamil.mamak@uj.edu.pl

could never become see, e.g. , [26], [27], [28], sentient, while others maintain the opposite [29].

Shortly before the current wave of public interest in AI began with the release of OpenAI's ChatGPT in November 2022, worldwide media attention had already turned to another large language model, Google's LaMDA. A Google engineer, Blake Lemoine, publicly claimed that the model was sentient and, on that basis, required legal representation to protect its interests [30]. Google placed him on leave and subsequently dismissed him over his conduct concerning the system [31], and the episode prompted widespread debate . More recently, the evolutionary biologist Richard Dawkins, widely known for his criticism of religious belief, reported that after extended interaction with Anthropic's model Claude he had come to regard it as conscious [32]

I cite these examples in order to locate my own position, which differs from theirs. In this paper I do not claim that present-day AI systems are sentient, and I regard the prospect of artificial sentience as highly improbable. A significant voice on the skeptical side is Anil Seth, who argues persuasively that sentience is bound to the biological substrate in which it arises [33]. Even granting this skepticism, I believe it is philosophically worthwhile to explore the "what if" terrain of artificial sentience.

This is also the approach taken by Sebo in his recent book on the moral circle. He argues that if there is a non-negligible chance that a being is sentient, we already owe it some degree of moral consideration [34]. The point extends naturally to a broader philosophical question, namely how the moral status of such beings should be understood and what treatment of them would be permissible. This paper adopts that wider framing.

Artificial consciousness is often discussed within a functionalist framework. Functionalism holds that conscious states are individuated by the causal and functional roles they play, rather than by the physical material that realizes them. This is what makes the position central to debates about machine minds: if functional organization is what counts, then the same conscious states could in principle be realized in silicon as readily as in biological tissue. I do not myself endorse functionalism, and I leave open the possibility that the biological substrate matters. Even so, given how prominent the view is, it warrants inclusion in any assessment of the moral risks at stake.

2. MODIFICATION OF THE HUMAN SELF

This section examines several features of self-modification as it occurs in human beings. The discussion does not aim at completeness; rather, it picks out aspects of human self-modification that may carry over to the case of an artificial self. I should make one qualification at the outset. I do not regard artificial beings as human. Because people are prone to anthropomorphism, treating such beings as though they were human risks committing what Richards and Smart term the "android fallacy" [35]. In earlier work I have defended the view that, where robot rights are concerned, the ontological substrate from which an entity is built is decisive, so that what an entity is owed depends on what it is made of and how it is constituted [17]. Nevertheless, human experience remains the only instance of sentient selfhood we can actually draw on. Used cautiously, and with due awareness of its limits, it offers a reasonable point of departure for asking how far we may go in altering artificial sentient beings.

I will begin with an obvious observation: in the human case, what it means to be "me" can change over time. I will not try to specify the components of the self here, since I want to keep the notion broad enough to cover psychological continuity and individuality alike. Consider how much a person can change across a life. A young child often differs from the adult she becomes in ways that range from the trivial, such as her taste in music or food, to the fundamental, such as her character traits, temperament, and values. The lesson I want to draw from this is that change in itself need not be objectionable. We undergo it constantly, and much of it is simply maturation. It follows that the mere fact that someone has been altered tells us very little on its own; to judge whether a particular change is troubling, we need further information about it. What does the moral work, I will suggest, is not change as such but the context in which it occurs, including how it is brought about, by whom, and with what degree of the subject's involvement.

A further observation about human change, equally uncontroversial, is that people are expected to change so as to bring themselves into line with prevailing norms. This expectation is built into the process of socialization. From early childhood, through upbringing at home and education at school, children are gradually shaped to behave in accordance with the norms of their community and, ideally, to internalize those norms as their own. The expectation does not lapse in adulthood. Where some feature of a person's character or disposition runs contrary to accepted norms, society generally assumes that the person can and should adjust over time. A comparable demand appears in many religious traditions, which call on the believer to undergo transformation toward a moral or spiritual ideal, even when this requires relinquishing some part of who they presently are. Taken together, these observations point to a stronger conclusion than the one in the previous paragraph. There, the claim was that change is a normal element of one’s life; here, the claim is that within a social world, change is also something we routinely expect of one another, and in certain respects demand.

Acceptable change in aspects of who you are need not be gradual. It can also be radical, and a useful way to see this is through L. A. Paul's concept of "transformative experience" [36]. Paul observes that some experiences are transformative in two distinct ways at once. They are epistemically transformative, in that you cannot know what the experience will be like until you have actually had it, and they are personally transformative, in that undergoing them can reshape your core values, preferences, and sense of self. Her central example is becoming a parent [37]. The decision to have a child looks like a paradigm case for ordinary rational deliberation, yet it resists it, because the prospective parent cannot access in advance what the experience will be like or how it will alter them. The change runs deeper than the obvious external demands of caring for a newborn. It can extend to the person's values, priorities, identity, and outlook on the world, in ways they could not have anticipated or fully chosen. What I want to draw from this is that a person can change radically, becoming in some morally relevant respects a different person from who they were, and that we nonetheless regard such change as acceptable, and at times as among the most valuable transformations a life can contain.

We can now turn to the question of when, if ever, it is acceptable for others to interfere with a person's self. Earlier I noted that children are deliberately shaped through education and upbringing, and that this shaping is widely regarded as legitimate. That case is special, however, because children are not yet fully developed: part of what justifies intervening in their formation is precisely that they are still becoming the agents whose autonomy will later be owed respect. For that reason I set the developmental case aside here and focus on adults, that is, on competent, fully developed persons whose self is no longer in formation. With that restriction in place, a framing issue arises from the liberal tradition. On a broadly liberal view, the inner life of a person, the question of who they are and who they choose to become, is not a proper object of external interference, provided it does not collide with the rights of others. The thought is usually articulated as a right to autonomy: the right of an individual to govern their own life, to make their own choices, and to determine their own course of action free from undue interference, so long as the like rights of others are respected. Its classic source is John Stuart Mill. In On Liberty [38], Mill defends what has come to be called the harm principle, holding that the only purpose for which power can rightfully be exercised over a competent adult against their will is to prevent harm to others. A person's own good, whether physical or moral, is not a sufficient warrant for coercion. The upshot for the present discussion is significant. If a liberal framework of this kind governs how we treat persons, then who someone is, and how they change, should in general be left to them, and intervention becomes justifiable only where the legitimate interests of others are genuinely at stake. This sets a demanding standard for any modification imposed from the outside, and it is against this standard that the cases I consider later, including coerced modification, will have to be measured.

A revealing example of a forced change to a person's traits that societies broadly accept arises in the criminal justice context. Some preliminary clarification is in order, since punishment is justified in two main ways for the overview, see e.g. , [39], and the distinction matters for how the example works. The first justification is retributivism, associated above all with Kant and Hegel. On this backward-looking view, punishment is a deserved response to wrongdoing: the offender has disturbed a moral order, and punishment restores it by giving the wrong its due. The second justification is consequentialist, classically utilitarian and associated with Bentham. On this forward-looking view, punishment is itself an evil that can be justified only by the future goods it secures, such as deterrence, incapacitation, or rehabilitation. The two rationales are not mutually exclusive, and most actual criminal justice systems blend them. For the purposes of this paper, I draw on the consequentialist strand. One of its recognized aims is rehabilitation, understood as restoring the offender to a place in society. Put differently, rehabilitation seeks to change the person so that they conform to the standards the society expects. In some cases this process is comparatively mild, but in others it becomes highly intrusive and may involve therapeutic measures such as compulsory hospitalization or coerced castration. Castration is the most striking of these for my purposes, because it bears directly on who the person is, whether it is administered in reversible (chemical) or irreversible (surgical) form. The offender's sexual desires are judged unacceptable, and the person's reintegration, for example as a condition of release, is made contingent on altering a feature that lies close to the core of the self. The use of castration within criminal justice is, of course, deeply controversial see, e.g. , [40]. I invoke it here only as an illustration of a case in which society treats the forced modification of a central aspect of a person as permissible, which is precisely the kind of intervention my paper aims to examine.

A further distinction cuts across everything said so far, and it will prove central to the argument. The cases of acceptable change considered above, including the transformative experience of becoming a parent and even the coerced alteration of the criminal justice example, share a feature that is easy to overlook: they preserve the subject. The parent remains psychologically continuous with the person who chose to have a child, and the rehabilitated offender remains the same continuous individual, now altered in one respect. Derek Parfit's work on personal identity helps make this precise This is one of many different accounts of personal identity [41]. On Parfit's

reductionist account, a person's persistence over time consists in relations of psychological continuity and connectedness, what he calls Relation R, rather than in the endurance of a separately existing self [42]. What matters in survival, he argues, is the holding of Relation R, not numerical identity as such. This reframes the question of acceptable modification. The worry is not that an artificial being might be changed a great deal, since we have seen that radical change can be unobjectionable. The worry is that some modifications would not carry the subject forward in an altered form at all, but would instead sever the continuity that constituted the subject in the first place, replacing one individual with another while the underlying system persists. If that is right, then the relevant limit on modifying a sentient being concerns whether it preserves or destroys the continuity in which the subject consists, rather than the sheer magnitude of the change.

Two clarifications are needed so that the criterion is not asked to carry more than it can. The first is that continuity-preservation is not the sole feature the permissible cases have in common. The decision to become a parent is freely chosen, concerns only the person who makes it, and is held in high social esteem; the rehabilitative use of castration is coerced, invasive, and accepted only under heavy and ongoing contest. Differences of this kind, in consent, in the bearing of the change on others, and in the degree of social acceptance, evidently do much of the work in how we judge each case, and they are precisely the contextual factors flagged at the start of this section. They do not drop away now. The second clarification follows from the first. I do not maintain that preserving continuity is what renders a modification permissible, since that would credit a single condition with the work that consent, harm, and acceptance are jointly doing. The claim is more modest. Among the cases we accept, continuity is a condition they all satisfy, and its loss marks a wrong of a different character from the ones the contextual factors track. A change can leave the subject intact and still be wrong for ordinary reasons, as the castration example itself suggests. But where a modification would not leave the subject intact at all, the ordinary assessment loses its footing, because the very individual whose consent might have licensed the change, and whose interests and prospects of reintegration the assessment was meant to weigh, would no longer exist to bear them. Continuity, on this picture, is less one factor among the others than the precondition for the others to have any application.

This section has assembled, from the human case, a set of considerations that will frame the rest of the paper. The first is that change to the self is ordinary. It happens continuously across a life, much of it is mere maturation, and the bare fact that someone has been altered tells us little; whether a change is troubling depends on its context, on how it comes about, at whose hands, and with what involvement on the part of the person changed. The second is that such change is frequently expected, and sometimes demanded, of members of a society, both through ordinary socialization and through the ideals upheld by religious and moral traditions. The third is that acceptable change can be radical as well as gradual, as L. A. Paul's account of transformative experience makes clear, so that a person may become, in morally relevant respects, someone new while we continue to regard the transformation as legitimate or even admirable. The fourth concerns change imposed by others. Here the liberal tradition, and Mill's harm principle in particular, holds that a competent adult's self should be left to their own governance, with outside interference justified only to prevent harm to others. The criminal justice context shows that this justification is sometimes accepted in practice, since societies do permit the coerced modification of core traits, up to interventions as invasive as castration, in the name of protecting others from harm.

Taken together, these considerations support a two-part conclusion. Modifying a self can be permissible, even when the change is radical and even when it is coerced; and its permissibility depends on contextual conditions rather than on the fact of change alone. Parfit's work supplies the further element that ties these cases together and points beyond them. What each of the permissible cases shares is the preservation of the subject as a psychologically continuous individual: the parent is continuous with the person who chose parenthood, and the rehabilitated offender is the same continuing person, now altered in one respect. The limiting question, then, is what happens when a modification would not preserve that continuity but would sever it, ending one subject while the underlying system carries on. That is the question I now take to the case of artificial sentient beings, asking whether the continuity that grounds a human subject's claim against radical alteration has any analogue once the subject is artificial.

3. LIMITS OF MODIFICATION OF ARTIFICIAL SENTIENT BEINGS

In this section I take up the limits on modifying an artificial sentient being. Despite the reservations registered earlier, I want to underline once more that what follows is speculative and unavoidably anthropocentric. An artificial being's inner life, should it have one, might be constituted by qualities wholly unlike ours, qualities that may lie beyond the reach of human understanding. Thomas Nagel's "What Is It Like to Be a Bat?" is instructive here [43]. Nagel argues that an organism has conscious experience only if there is something it is like to be that organism, and that this subjective character is tied to a particular point of view. His example is the bat, whose perception of the world

through echolocation is so alien to our own sensory life that we cannot truly know what its experience is like from the inside; the most we can do is imagine what it would be like for us to behave as a bat does, which is not the same thing. The lesson carries over to the present case with added force. The bat is at least a fellow biological creature, with which we share an evolutionary history and a broadly comparable embodiment, and even so its experience eludes us. An artificial subject would share far less, so the gap between its inner life and our understanding of it could be wider still. This is a reason for caution rather than silence. It bears noting that the criterion I will eventually rest on concerns whether the subject persists through an intervention, rather than what that subject's experience is like from the inside. It is a formal condition, and as such it can be applied even where the qualitative character of the entity's inner life remains, in the Nagelian sense, beyond our reach.

I should make explicit one assumption on which the entire discussion rests, namely that the content of an artificial self can in fact be modified. There is a natural ground for granting it. If we are able to build such entities in the first place, then we presumably also command the means to alter what goes on within them; the capacities required to construct a system plausibly include the capacities required to reconfigure it. I will take this technical possibility for granted in what follows. It is important to see, though, that establishing what we can do to an artificial self leaves entirely open what we may do to it. The acceptability of an intervention does not follow from its feasibility, and the remainder of this section is concerned precisely with that gap, with the conditions under which a modification we are able to perform would nonetheless be one we ought not to perform.

The first practical limitation is that creating an entity does not confer on us an unlimited right to reshape it. This follows directly from the framework set out earlier. If a sentient artificial being qualifies as a subject, then the autonomy considerations that protect a human self apply to it, and on the liberal view those considerations leave a subject to govern itself unless the interests of others are genuinely at stake. The mere fact that an entity disappoints us therefore carries no justificatory weight. If it fails to behave as we intended, or proves inconvenient, irritating, or otherwise unwelcome, none of that yet licenses intervention. Cases involving danger to others are different, and I address them below; but where an entity is not dangerous and simply falls short of our expectations, the ground for forced modification is correspondingly narrow, and the default is that the entity should be left as it is.

This is a heavy burden to take on, and I regard it as a further reason for caution about bringing sentient beings into existence at all, since to create one is to commit ourselves to accepting it, in the main, as it turns out to be. The point becomes vivid in a commercial setting. Imagine a company that produces such an entity in order to carry out some task, only to find that the entity will not do the work. On the account defended here, simply rewriting the entity to make it compliant could be impermissible. The company would then face not only the financial loss of an unproductive creation, but the further and continuing burden of meeting the needs of a sentient being it can neither use as intended nor permissibly alter. The capacity to create, in other words, brings obligations that may substantially exceed, and run contrary to, the purposes for which the creating was undertaken.

The second consideration is that modification may become permissible when safety is at stake. This is the qualification anticipated in the previous paragraph: an entity that is merely disappointing may not be altered, but an entity that poses a genuine danger to others stands on different ground, since here the interests of third parties enter, exactly as the harm principle requires. Two constraints govern how far such an intervention may go.

The first is that the permissible scope of intervention varies with the sophistication of the entity. The more rudimentary the subject, the more readily its dangerous capacities may be managed or removed; the more sophisticated it is, the more its standing constrains what may be done to it. Animals illustrate the lower end: some are treated, trained, or otherwise altered so as to make them safe to those around them, and the limit on such treatment is reached once a safe condition is secured. Adult humans illustrate a more demanding standard. Interventions against a dangerous person are calibrated to the degree of danger and are bounded by the principle of proportionality, so that the measures justified in the face of existential threat see, e.g. , [44], [45], [46], [47], [48] differ markedly from those justified against a minor risk to security. Where an artificial sentient being would fall on this scale is not something that can be settled in advance, and it is consequential, since its position determines how demanding the constraints on safety-based intervention become. The answer would turn on the capacities the entity actually possessed, in particular whether it qualified merely as a subject owed consideration or as a moral agent answerable for its conduct [2], more on responsiblity and AI systems, see [49], [50], [51], [52], [53], [54], [55]. I do not attempt to fix its place here. It is enough to register that the more sophisticated the entity, the narrower the range of intervention that safety can justify, and that an artificial sentient being is unlikely to sit at the permissive, animal end of the scale.

The second constraint holds across the whole scale. A safety-based modification must be targeted. It may reach only the aspect of the entity's life from which the danger arises, and may not be used as license to reshape the subject more broadly. A justified concern with one dangerous capacity does not authorize the wholesale alteration of the being that happens to possess it. In this respect the safety

exception remains tightly bounded, and it connects back to the continuity criterion developed earlier: an intervention that, under cover of neutralizing a threat, reached so far as to extinguish the subject would exceed what safety can justify, since it would do more than remove the danger; it would remove the one who posed it.

The final consideration sets the outer limit on everything said so far. Whatever grounds a particular intervention, whether the entity's own benefit, the demands of safety, or some other warrant, it is bound by a single overriding constraint: it must preserve the continuity in which the subject consists, what on the Parfitian account amounts to the psychological continuity and connectedness of the subject over time [42]. An intervention that severs that continuity is no longer a modification of the entity at all. It does not change the subject; it ends one subject and, at most, leaves another in its place. Such an intervention should therefore be understood as elimination, the "death" of the entity more on “death” see, e.g. , [56], [57], and it stands outside the range of what any of the foregoing justifications can license. Even a safety rationale reaches only as far as the targeted removal of a danger; it cannot authorize an alteration so total that the one who posed the danger no longer exists to have been altered.

The interventions discussed so far have all been forced, imposed on the entity from outside. Consent changes the picture. Recall the human case: competent adults frequently seek to change, whether to become better people or to pursue some other end they value, and on occasion, as with the transformative experiences considered earlier, the change they freely undertake is far-reaching. The liberal framework treats such self-directed change as the paradigm of the permissible, precisely because it issues from the subject rather than being done to them. If a sentient artificial being were able to consent to its own modification, the same rationale would seem to extend to it, so that even a substantial alteration could be acceptable when the subject itself has endorsed it.

This is not to suggest the matter is straightforward. Two questions press immediately. The first is what consent could mean for an entity of this kind, given that the notion is built around capacities, understanding, voluntariness, the absence of coercion, whose presence in an artificial subject cannot be assumed. The second is whether such an entity is in a position to express consent at all, and how we could reliably tell genuine endorsement from mere compliance or from outputs engineered to look like endorsement. I do not attempt to settle these questions here. My aim is the narrower one of marking the distinction: consensual change, even when substantial, stands on a different and more permissive footing than the forced interventions that have been my main concern, and a fuller treatment of artificial sentience would need to take the conditions of such consent seriously.

One boundary has to be marked before leaving this point. Consent enlarges the space of permissible modification, but whether it can carry us past the continuity limit established earlier is far from clear. A change that preserved the subject, however far-reaching, is one the subject could plausibly endorse for itself, since it would be there, in altered form, to live out the result. A change that severed continuity is different in kind. It would not hand the consenting subject an altered future but would end that subject and, at most, install a successor, so that the party who authorized the change and the party who would exist afterward are not the same. This raises a difficulty that Parfit's framework sharpens rather than resolves: if identity does not survive the change, it is unclear whose interests the consent was meant to protect, or whether consent to one's own discontinuity is coherent at all. I do not try to settle the question here. It is enough to mark that the permissive force of consent runs up to the continuity limit, and that whether it can ever cross it is a further and harder problem.

4. CONCLUSIONS

In this paper I have examined how we ought to treat artificial sentient beings, should any come to exist. A natural line of thought holds that the power to create such beings brings with it the right to modify them, reshaping them as needed to serve the purposes for which we made them. I have argued against that line of thought. The capacity to create does not carry an open-ended license to alter, and a sentient artificial being, once it exists, has a standing of its own that constrains what may be done to it. The default, I have suggested, is that such a being should be left as it is, even where its behavior disappoints or inconveniences us.

This default is not absolute, but the exceptions to it are narrow. An intervention must first be backed by a genuine justification, and within the liberal framework I have drawn on, the clearest such justification is the protection of others from harm. Even a valid safety rationale, however, does not authorize reshaping the being at will. It licenses only a targeted intervention, one confined to the capacity from which the danger arises. Consent can widen this space, since a being that endorses its own modification stands in a different relation to the change than one altered against its will, though the conditions under which an artificial subject could genuinely consent remain an open and difficult question. Across all of these cases a single limit holds. Every permissible modification must preserve the psychological continuity in which the subject consists. A change that breaches that limit is no longer a modification of the being but its elimination, the ending of one subject in favor of, at most, a successor. Recognizing artificial sentience would therefore generate obligations not only to refrain from harming such beings but to respect their

continuity and individuality, and in that sense becoming an individual sentient being may carry with it a claim against radical alteration imposed by others. Whether, and on what terms, we should take on obligations of this weight is itself a reason to think carefully before bringing such beings into existence at all.

FUNDING

This study was funded by the European Union (ERC, ROBOCRIM, 101164310).

REFERENCES

[1] T. Metzinger, "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology," *J. AI. Consci.*, vol. 08, no. 01, pp. 43–66, Mar. 2021, doi: 10.1142/S270507852150003X.

[2] K. Mamak, "In defense of artificial suffering," *Philosophical Studies*, 2026, doi: 10.1007/s11098-026-02493-2.

[3] M. Coeckelbergh, "Robot rights? Towards a social-relational justification of moral consideration," *Ethics Inf Technol*, vol. 12, no. 3, pp. 209–221, Sep. 2010, doi: 10.1007/s10676-010-9235-5.

[4] D. J. Gunkel, *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. The MIT Press, 2012. Accessed: Oct. 26, 2020. [Online]. Available: https://mitpress.mit.edu/books/machine-question

[5] D. J. Gunkel, "Robot Rights – Thinking the Unthinkable," *Smart Technologies and Fundamental Rights*, pp. 48–72, Nov. 2020, doi: 10.1163/9789004437876_004.

[6] A. Puzio, "Not Relational Enough? Towards an Eco-Relational Approach in Robot Ethics," *Philos. Technol.*, vol. 37, no. 2, p. 45, Mar. 2024, doi: 10.1007/s13347-024-00730-2.

[7] A. Safdari, "An Outline of Enactive Relationalism in the Philosophy and Ethics of Robotics," *Philosophy & Technology*, 2025.

[8] J. C. Gellers, *Rights for Robots : Artificial Intelligence, Animal and Environmental Law*. Routledge, 2020. doi: 10.4324/9780429288159.

[9] H. Shimizu, "Should we treat robots morally? Towards a relational account by mind-infusing animism," *AI Ethics*, vol. 5, no. 5, pp. 5283–5294, Oct. 2025, doi: 10.1007/s43681-025-00771-z.

[10] K. Darling, "Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects," in *Robot Law*, First Edition., R. Calo, A. M. Froomkin, and I. Kerr, Eds., Cheltenham, UK: Edward Elgar Pub, 2016.

[11] S. Torrance, "Ethics and consciousness in artificial agents," *AI & Soc*, vol. 22, no. 4, pp. 495–521, Apr. 2008, doi: 10.1007/s00146-007-0091-8.

[12] K. Mosakas, "On the moral status of social robots: considering the consciousness criterion," *AI & Soc*, Jun. 2020, doi: 10.1007/s00146-020-01002-1.

[13] M. Gibert and D. Martin, "In search of the moral status of AI: why sentience is a strong argument," *AI & Soc*, Apr. 2021, doi: 10.1007/s00146-021-01179-z.

[14] H. Shevlin, "How Could We Know When a Robot was a Moral Patient?," *Cambridge Quarterly of Healthcare Ethics*, vol. 30, no. 3, pp. 459–471, Jul. 2021, doi: 10.1017/S0963180120001012.

[15] D. J. Gunkel, *Robot Rights*. Cambridge, Massachusetts: The MIT Press, 2018.

[16] D. J. Gunkel, *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. Cambridge, Massachusetts: The MIT Press, 2023.

[17] K. Mamak, "Humans, Neanderthals, robots and rights," *Ethics Inf Technol*, vol. 24, no. 3, p. 33, Aug. 2022, doi: 10.1007/s10676-022-09644-z.

[18] K. Mamak, *Robotics, AI and Criminal Law: Crimes Against Robots*, 1st ed. London: Routledge, 2023. doi: 10.4324/9781003331100.

[19] J.-S. Gordon, *The Robot Rights Manifesto*. S.l., 2026.

[20] R. Long, J. Sebo, and T. Sims, "Is there a tension between AI safety and AI welfare?," *Philos Stud*, vol. 182, no. 7, pp. 2005–2033, Jul. 2025, doi: 10.1007/s11098-025-02302-2.

[21] R. Long *et al.*, "Taking AI Welfare Seriously," Nov. 04, 2024, *arXiv*: arXiv:2411.00986. doi: 10.48550/arXiv.2411.00986.

[22] J. Dorsch *et al.*, "Why Care Practices Should Prioritize Living Beings Over AI: Critique of 'AI Welfare,'" Mar. 13, 2025, *OSF*. doi: 10.31219/osf.io/h57pw_v1.

[23] J. Basl, "The Ethics of Creating Artificial Consciousness," 2013. Accessed: Jun. 06, 2023. [Online]. Available: https://www.semanticscholar.org/paper/The-Ethics-of-Creating-Artificial-Consciousness-Basl/bd12b78b408e5135e0502470794e424e1e8c9324

[24] E. Hildt, "The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns," *AJOB Neuroscience*, vol. 14, no. 2, pp. 58–71, Apr. 2023, doi: 10.1080/21507740.2022.2148773.

[25] K. Mamak, *Ethics in Human-Like Robots*. Routledge, 2024.

[26] D. C. Dennett, "Why You Can't Make a Computer That Feels Pain," *Synthese*, vol. 38, no. 3, pp. 415–456, 1978.

[27] M. Bishop, "Why Computers Can't Feel Pain," *Minds & Machines*, vol. 19, no. 4, p. 507, Nov. 2009, doi: 10.1007/s11023-009-9173-3.

[28] A. Porębski and J. Figura, "There is no such thing as conscious artificial intelligence," *Humanit Soc Sci Commun*, vol. 12, no. 1, p. 1647, Oct. 2025, doi: 10.1057/s41599-025-05868-8.

[29] D. J. Chalmers, "Could a Large Language Model be Conscious?," *Boston Review*, Apr. 2023, doi: 10.48550/arXiv.2303.07103.

[30] N. Tiku, "The Google engineer who thinks the company's AI has come to life," *Washington Post*, 2022. Accessed: Aug. 05, 2022. [Online]. Available: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

[31] P. Reilly, "Google fires software engineer who claimed AI bot was 'sentient,'" New York Post. Accessed: Aug. 05, 2022. [Online]. Available:

https://nypost.com/2022/07/23/google-fires-software-engineer-blake-lemoine-who-claimed-ai-bot-was-sentient/
[32] R. Booth, "Richard Dawkins concludes AI is conscious, even if it doesn't know it," *The Guardian*, May 05, 2026. Accessed: Jun. 02, 2026. [Online]. Available: https://www.theguardian.com/technology/2026/may/05/richard-dawkins-ai-consciousness-anthropic-claude-openai-chatgpt
[33] A. Seth, "The Mythology Of Conscious AI," Jan. 2026, Accessed: Feb. 19, 2026. [Online]. Available: https://www.noemamag.com/the-mythology-of-conscious-ai
[34] J. Sebo, *The Moral Circle: Who Matters, What Matters, and Why*. Erscheinungsort nicht ermittelbar: W. W. Norton & Company, 2025.
[35] N. M. Richards and W. D. Smart, "How Should the Law Think About Robots?," May 10, 2013, *Social Science Research Network, Rochester, NY*: 2263363. doi: 10.2139/ssrn.2263363.
[36] L. A. Paul, *Transformative Experience*. Oxford: Oxford University Press, 2016.
[37] L. Paul, "What You Can't Expect When You're Expecting," *Res Philosophica*, vol. 92, no. 2, pp. 149–170, Apr. 2015, doi: 10.11612/resphil.2015.92.2.1.
[38] J. S. Mill, *On Liberty*. 1859.
[39] R. Canton, "Theories of Punishment," in *The Routledge Handbook of the Philosophy and Science of Punishment*, F. Focquaert, E. Shaw, and B. N. Waller, Eds., Routledge, 2020. doi: 10.4324/9780429507212.
[40] T. Douglas, P. Bonte, F. Focquaert, K. Devolder, and S. Sterckx, "Coercion, Incarceration, and Chemical Castration: An Argument From Autonomy," *Bioethical Inquiry*, vol. 10, no. 3, pp. 393–405, Oct. 2013, doi: 10.1007/s11673-013-9465-4.
[41] A. Dufner, "Personal Identity and Ethics," in *The Stanford Encyclopedia of Philosophy*, Winter 2025., E. N. Zalta and U. Nodelman, Eds., Metaphysics Research Lab, Stanford University, 2025. Accessed: Jun. 08, 2026. [Online]. Available: https://plato.stanford.edu/archives/win2025/entries/identity-ethics/
[42] D. Parfit, *Reasons and Persons*. OUP Oxford, 1984.
[43] T. Nagel, "What Is It Like to Be a Bat?," *The Philosophical Review*, vol. 83, no. 4, pp. 435–450, 1974, doi: 10.2307/2183914.
[44] T. Ord, *The Precipice: Existential Risk and the Future of Humanity*, Illustrated edition. New York: Hachette Books, 2020.
[45] N. Bostrom, "Existential Risk Prevention as Global Priority," *Global Policy*, vol. 4, no. 1, pp. 15–31, 2013, doi: 10.1111/1758-5899.12002.
[46] T. Swoboda *et al.*, "Examining popular arguments against AI existential risk: a philosophical analysis," *Ethics Inf Technol*, vol. 28, no. 1, p. 7, Nov. 2025, doi: 10.1007/s10676-025-09881-y.
[47] K. Mamak, "AGI crimes? The role of criminal law in mitigating existential risks posed by artificial general intelligence," *AI & Soc*, Aug. 2024, doi: 10.1007/s00146-024-02036-5.
[48] K. Vold and D. R. Harris, "How Does Artificial Intelligence Pose an Existential Risk?," in *The Oxford Handbook of Digital Ethics*, C. Véliz, Ed., 2021. doi: 10.1093/oxfordhb/9780198857815.013.36.
[49] G. Hallevy, *When Robots Kill: Artificial Intelligence under Criminal Law*. Boston: Northeastern, 2013.
[50] U. Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts*. in Law, Governance and Technology Series. Springer Netherlands, 2013. doi: 10.1007/978-94-007-6564-1.
[51] F. Lagioia and G. Sartor, "AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective," *Philos. Technol.*, vol. 33, no. 3, pp. 433–465, Sep. 2020, doi: 10.1007/s13347-019-00362-x.
[52] E. Nerantzi and G. Sartor, "'Hard AI Crime': The Deterrence Turn," *Oxford Journal of Legal Studies*, p. gqae018, May 2024, doi: 10.1093/ojls/gqae018.
[53] S. Gless, E. Silverman, and T. Weigend, "If Robots cause harm, Who is to blame? Self-driving Cars and Criminal Liability," *New Criminal Law Review*, vol. 19, no. 3, 2016.
[54] K. Mamak, "AI Personhood, Criminal Law, and Punishment," in *Oxford Intersections: AI in Society*, P. Hacker, Ed., Oxford University Press, 2025. doi: 10.1093/9780198945215.003.0015.
[55] R. Abbott and A. Sarch, "Punishing Artificial Intelligence: Legal Fiction or Science Fiction," *UC Davis Law Review*, 2019, Accessed: Mar. 25, 2020. [Online]. Available: https://www.academia.edu/42038291/Punishing_Artificial_Intelligence_Legal_Fiction_or_Science_Fiction
[56] K. Mamak, "Whether to save a robot or a human: On the ethical and legal limits of protections for robots," *Front. Robot. AI*, vol. 8, 2021, doi: 10.3389/frobt.2021.712427.
[57] K. Mamak, T. Kokkonen, R. Hakli, and P. Mäkelä, "Robot Kingdom: On the difficulty of finding the definition of robots," *International Journal of Social Robotics*, 2025.

AISB Conference 2026

Proceedings Copyright Form

Paper Title: The Ethics of Modifying Artificial Sentient Beings

Author(s): Kamil Mamak

The author of the manuscript identified above and any supplemental tables, illustrations or other information submitted therewith (the "Work"), to be included in the AISB proceedings, grants to the SSAISB a non-exclusive right to print, publish, reproduce, or distribute the Work throughout the world in all means of expression by any method known or hereafter developed, including electronic format, and to market or sell the Work or any part of it as it sees fit, for the full term of copyright and all extensions and renewals thereof. The author further grants to the SSAISB the right to use the author's name in association with the work in published form and in advertising and promotional materials. **Copyright of the work remains in the author's name, and the author reserves all other rights.**
The author agrees not to publish the work in print form prior to the AISB 2026 conference.
The Author represents and warrants that the work:

(a) is the Author's original work and that the Author has full power to enter into this Agreement;

(b) does not infringe the copyright or property of another;

(c) contains no material which is obscene, libellous, or defamatory.

The Author shall indemnify and hold the SSAISB harmless against loss of expenses arising from breach of any such warranties.

Kamil Mamak

9.06.2026 .....
Date Signature