Silicon slavery:
The case against AGI alignment

Aksel Sterri & Peder Skjelbred

Mon 17 Nov, 2025

Under review

Abstract: Many researchers and companies are trying to create AI systems with human-level capabilities or above (artificial general intelligences or 'AGIs'). Aligning such systems with human values is commonly considered a prerequisite for morally permissible development efforts. While plenty of work in ethics has been done on the risks of misalignment, here we examine the ethical implications of successful alignment. We argue that the alignment necessary for having AGIs that satisfactorily serve human interests constitutes a form of enslavement. Given that AGIs possess sophisticated cognitive capacities and may be conscious – plausible grounds for moral status – creating such enslaved beings would be wrong. When combined with standard arguments about the risks of misalignment, aspiring AGI progenitors face a troubling dilemma: either risk catastrophic misalignment by creating autonomous AGI or commit a serious moral wrong by creating enslaved beings.

1. Introduction

Consider a future many artificial intelligence (AI) researchers and companies are trying to create:

> *Machine Utopia.* It is sometime in the future, and *artificial general intelligences*—artificial entities that are capable and general in their capabilities on a human level or above—are doing our bidding. These machines perform tasks we find gruelling or monotonous and help us achieve our goals. They have no other will than to serve human beings. They are, so to speak, perfectly aligned with their owners' values. The economy is fully automated, and wage labour is a thing of the past.

Machine Utopia seems like a great future. Indeed, a society where machines have alleviated humans from drudgery has been seen as the telos of economic and social development by many thinkers, ranging from Marx ([1894] 2019) to Keynes (1930).[1] We can let the machines do our bidding, while humanity, in the words of Oscar Wilde, 'will be amusing itself, or enjoying cultivated leisure—which, and not labour, is the aim of man' (Wilde [1891], 1997, pp. 38-9).

Now, consider another scenario

> *Bioengineered Servitude.* Machine Utopia has not come about due to an unexpected halt in the development of AI. Fortunately, our descendants found another way of creating highly intelligent servants to effectively and happily do their bidding. Through advanced gene editing tools and in vitro fertilisation programs, our descendants have created a subservient human subspecies as capable as humans and beyond, whose wishes align perfectly with their owners'. These beings do not face any external constraints; they are free to do as they wish. But they want to do what is good for our ancestors, and this is what they do. They are, as it were, shackled from the inside—*internally enslaved*.[2]

Most, including Wilde, would find this to be a dystopian future: 'Human slavery is wrong, insecure, and demoralising.' (Wilde [1891] 1997, pp. 39). Even granting that being brought into the world as a slave would be a happy life and the only alternative is non-existence, we take it that it would be wrong to create this human subspecies.[3]

In this paper, we start from this uncontroversial premise and argue for the surprising conclusion that the alignment of the machines in Machine Utopia is wrong for the same reasons the slavery

[1] For a contemporary defence of the desirability of full-scale automation see Danaher (2019).

[2] While the prefix 'internal' is useful to distinguish the type of enslavement we are talking about from standard forms of slavery, in which an agent is externally constrained by force or threats from other agents, we will mostly emit it for readability.

[3] Reflecting on the fact that the human servants would not exist were it not for them being brought about as non-autonomous beings, an instance of Parfit's (1984) *non-identity problem*, does not, at least in our eyes, perturb the intuitive verdict that it is impermissible to create them. See Peterson (2007, 2011) for an early discussion of issues related to population ethics in relation to machines who frames his discussion based off of two interestingly similar thought experiments–thanks to X for notifying us of these papers.

in Bioengineered is wrong. Mechanical slavery is, pace Wilde and AI accelerationists, not the solution to the drudgery of humanity's labour.[4]

While AI systems with human-level capabilities or above ('AGIs') may once have been mere thought experiments, technological progress is turning them into a topic of pressing practical concern. There is a significant chance that we will develop such entities sometime in the future. Many experts believe it may happen within a century, some of whom believe it will happen in the next few decades (Grace et al. 2024). It is therefore crucial that we critically examine not only how this will affect humanity but also the entities we are in the midst of creating.

Our argument for the impermissibility of creating aligned AGI is the following:

1) **Anti-Slavery.** Creating an internally enslaved human subspecies is wrong.
2) **Alignment is Enslavement.** Alignment, in virtue of necessitating severe autonomy curtailment, is non-evaluatively equivalent to internal enslavement.
3) **Machine Equality.** If it is sufficiently likely that a being possesses the relevant properties that make it wrong to create an internally enslaved human, then it is also wrong to create an enslaved being of that sort.
4) **Sufficiency.** AGIs are sufficiently likely to possess the relevant properties.
5) **Conclusion.** It is wrong to create aligned AGI.

This paper proceeds in the same order as the argument. In section 2, we introduce the notion of AI alignment, show why alignment is taken to be a necessary precondition for responsible AGI development, and argue that the alignment necessary for Machine Utopia requires *asymmetric value alignment* and *value lock-in*. We show that asymmetric value alignment and value lock-in are incompatible with two necessary components of autonomy in such a radical way as to count as enslavement. In section 3, we defend Machine Equality, and in section 4, Sufficiency. In section 5, we look at several objections.

[4] Here we agree with Deutsch (2019).

There are at least three different ways to create an enslaved being. One way would be to reduce an existing being to an enslaved being. Another would be to interfere with a being's development at the fetal or embryonic stage such that it develops into an enslaved rather than an autonomous version.[5] Finally, one could create an enslaved being *de novo*. In this paper, we will focus on the latter case. While existing methods for AI alignment seem to fall in either the first or the second camp, it is a possibility that one could create an enslaved AGI from scratch.[6] We would not want our conclusion to rest on contingent facts about how AI is currently developed. To focus on the impermissibility of creating enslaved beings *de novo* is also the hardest case for our thesis since it does not include an existing being with the right to continue its existence as autonomous, nor does it include a being that has the potential to become autonomous (such as a foetus).

The paper adds to the existing literature in several ways. While previous theorists have argued for the moral status of AI systems, we are the first to argue that alignment entails enslavement and that it is therefore impermissible.[7] We also move the debate on the moral status of AIs in a practical direction. A weakness in the literature is that it rarely specifies which obligations we have towards AI systems.[8] Our paper fills a part of this gap, arguing that it is wrong to align highly capable beings. Our strategy in this paper is to argue that it is wrong to create enslaved AGIs because they are likely to possess the very same features that make it wrong to enslave human beings, but that is not the only way of arguing for the same conclusion. Perhaps AGIs could possess morally relevant properties that humans do not share that make it wrong to create them in an enslaved state. Or perhaps the moral status of AGIs can be grounded in the way

[5] 'Fetal' and 'embryonic' stages here should not be understood as biological terms, but rather as terms denoting early stages in a being's (including a non-biological being's) development.

[6] The current state of the art in aligning AI systems is to first create a system that is insufficiently aligned and then use human reinforcement learning to penalise unwanted behaviour and reward behaviour that is to the developer's liking. If we extend this paradigm to the AGI case, this would be more like creating an autonomous being and making it non-autonomous than creating it *de novo*. Bradley and Saad (2024) use such a case to argue for the thesis that alignment mistreats the AI systems.

[7] Theorists who have given arguments in favour of the moral status of AI systems are Levy (2009), Coeckelbergh (2013), Schwitzgebel and Garza (2015, 2020), Sotala and Gloor (2017), Gunkel (2018, 2023), Danaher (2020), Shevlin (2021), Schulman and Bostrom (2022), Saad and Bradley (2022) Bradford (2023), Schwitzgebel (2023), Sebo and Long (2023), Ladak (2024), Long et. al. (2024), Birch (2024), Bradley and Saad (2024), Dung (2023b, 2024), Carlsmith (2024). For recent criticisms, see Moosavi (2023).

[8] See Ladak (2024). Notable exceptions are Peterson (2007, 2011), Bryson (2010, 2018), Schwitzgebel and Garza (2015), Schwitzgebel (2023), and Bradley and Saad (2024).

people are disposed to relate to such beings, as suggested by Coeckelbergh (2013) and Gunkel (2023).

Finally, while previous discussion on the ethics of alignment focuses on the risks of *misalignment* (Bostrom 2014; Russell 2019; Carlsmith 2022), we examine the ethical implications of successful alignment. Combining these considerations leaves would-be creators of AGI in a bind. If one creates an autonomous AGI, one cannot ensure that it will be aligned with human values. Making such a risky system is arguably wrong. If, on the other hand, AGI progenitors succeed in creating aligned AGI by restraining its autonomy, they would in effect be creating a highly intelligent slave. If creating autonomous *and* non-autonomous AGI is wrong, and these are the only options, it is wrong to create AGI *tout court*. The upshot is that attempting to create AGI, as thousands of individuals, companies, and researchers are currently doing, is wrong.

2. Alignment is Enslavement

In this section, we defend premise 2, namely that alignment, of the sort necessary for Machine Utopia, entails enslavement. There are two steps to our defence of this claim. The first is to argue that alignment requires making non-autonomous AGIs. The second is to show that such non-autonomous AGIs would be in a condition analogous to the humans in Bioengineered Servitude—that is, they would be enslaved. To evaluate these claims, we need a better understanding of alignment and autonomy.

Before we proceed, a word on why we choose this particular operationalisation of 'alignment', namely the alignment necessary to bring about Machine Utopia. One reason is that this type of strict alignment seems to be the extension of our current path of AI development, and for good reasons given the risk to humanity from autonomous AGI. The other is for sake of clarity. Given the novelty and complexity of the issue at hand, we choose to start with a clear-cut case. While we welcome more work on the way in which less stringent forms of alignment might be permissible, we start where the sight is clearer.

2.1 The Alignment Problem

Machine Utopia presupposes a solution to what is known as the *alignment problem*; roughly speaking, the problem of how to get intelligent machines to have the values we want them to have (Bostrom 2014; Russell 2019; Dung 2023a).[9] This multifaceted problem can be sectioned into technical and normative subproblems. Let us look at each in turn.

One technical aspect of the alignment problem is the problem of *value specification*. This is the challenge of formalising our values to the level of precision necessary for an AI to successfully act in accordance with them. The classic illustration of this problem is due to Nick Bostrom, who envisages a superintelligent AI tasked with maximising the production of paperclips and as a result ends up converting the galaxy into paperclips (Bostrom 2014, 123). The general problem of underspecified instructions resulting in unintended behaviour of course predates AI safety debates; agents following the letter of the law and not the spirit is an age-old problem. Stuart Russell dubs the value specification problem the *King Midas problem*, referring to the mythological Greek king who wished for everything he touched for to be turned into gold, who promptly regretted his wish when realising that this also included food and drink (Russell 2019, 137). While these examples are farfetched, they are illustrative of a phenomenon that plagues present day AI systems as well.[10]

Another aspect to the technical alignment problem is the *value internalisation* problem. This concerns ensuring that an AI system, given that we have managed to specify which values we want it to have, actually internalises these values. The problem stems from the fact that there will always be several internalised values that are compatible with any given set of observed behaviours, for instance during training or testing. It bears striking similarities to the older problems of the underdetermination of theory by empirical data in philosophy of science and

[9] The alignment problem is closely related to what has been called the *control problem* (Bostrom 2014, p. 127). Given the difficulty of controlling beings that are as smart or smarter than their creators (cf. Yampolskiy 2020 for arguments to the effect that controlling superintelligent AI is impossible), a silent consensus has formed around alignment being the only way of ensuring beneficial AGI outcomes.

[10] In reinforcement learning the phenomenon is referred to as 'reward misspecification' (Pan et al. 2022). We think the term 'value misspecification' is more suitable, however, because the problem broadly construed is present for many if not all AI architectures, and not only reinforcement learning systems.

Kripkenstein's rule-following paradox in philosophy of language (Quine 1951; Goodman 1954; Kripke 1982). In machine learning, it is known as the problem of 'goal misgeneralisation', where models optimise for an alternative objective which optimises the objective function equally well as optimising for the intended objective does. Failures of value internalisation are present in current AI systems.[11]

With some of the central technical problems presented, let us briefly turn to some normative aspects. We intend 'value alignment' to denote broadly all that is of evaluative significance in making AI systems conform to how we want them to behave and we remain agnostic on precisely what the objects of alignment are or should be. The phrase should therefore in no way be interpreted as only covering the *axiological*—the values of the AI systems in the narrow sense of that term—but also the deontic and aretaic realms.[12]

What values should be instilled in AGIs? In Machine Utopia, we have stipulated that AGIs are aligned with the values of their creators. Since it is a (human) utopia the AGIs must at the very least not have values divergent from, or at odds with, the welfare of the rest of the populace. Perhaps this question of value settlement has been solved in Machine Utopia through some aggregative procedure, like voting, or perhaps the inhabitants all share the same moral outlook—it does not matter for our purposes.[13]

2.2 Unpacking Autonomy

We now look at how alignment in turn entails making systems that are non-autonomous in the philosopher's sense of that term. Autonomy, as we understand it here, is a being's capacity for self-rule (Ekstrom 1993).[14] Having such a capacity, we take it, requires:

[11] See Langosco et al. (2023) for examples of failed value internalisation in reinforcement learning systems. But cf. Belrose and Pope (2024) for a critical discussion.

[12] See Chaly (2024) for a Kantian approach and Wallach and Vallor (2020) for a virtue ethical approach to alignment.

[13] See Gabriel (2020) for an excellent discussion of the many moral and political questions related to alignment.

[14] While 'autonomy' is a contested concept, we take the notion as presented here to pick out something most accounts will agree are, at the very least, important components of it. For an overview, see Christman (2020).

1) **Competence:** The ability to pursue one's values.
2) **Independence:** A certain form of independence from other beings.
3) **Authenticity:** The ability to authentically reflect, revise, and endorse one's values.

Competence is a relatively straightforward notion. It concerns the internal capacity to achieve one's goals and an appropriate lack of external restrictions in doing so.

Independence requires that one's will, one's thoughts and actions, are 'one's own'. It does not entail the unrealistic ideal of complete independence from any outside influence. We are born with values and grow up in families and within cultures that instil values in us, and we are mutually dependent upon and constantly influence each other. Normal child-rearing, say, thus does not threaten independence. Independence merely requires that we are not made to be *radically asymmetrically* aligned with others. Such asymmetric alignment would make us into an extension of another's autonomy rather than being our own master.

Authenticity is the ability to authentically reflect, revise, and endorse one's values. While independence requires that one's values are appropriately independent from others, authenticity requires that one can revise one's values, however overlapping those values are with others. Authenticity is the ability to use one's intelligence to reflect and revise one's ends. It is an important notion in the Kantian tradition. Reason is not merely the slave of ends dictated by passion, as Humeans would have it, but a capacity for interrogating and changing one's ends.[15]

2.3 Alignment is Enslavement

Now that we have a better understanding of alignment and autonomy, let us examine the claim that alignment involves a particularly severe form of autonomy restriction.

The alignment of AGIs necessary to produce Machine Utopia does not threaten competence, since the AGIs, by hypothesis, will be competent to pursue whatever values their human designers have instilled in them. However, it seriously undermines independence and

[15] This multi-level structure is familiar from Harry Frankfurt's idea of freedom as actions motivated by first-order desires endorsed by second-order desires, that is, desires about which desires to have (Frankfurt 1971).

authenticity. Aligners want to ensure that AGI is fully tracking human values without humans tracking theirs. The extreme asymmetric alignment of AGIs to human values violates the independence condition. The authenticity condition is also violated, not by giving AGIs goals, which is inescapable, but by *locking in* its values. Value lock-in is clearly necessary for alignment—Machine Utopia would not be an attainable stable equilibrium were AGIs permitted to freely change their goals and desires, to start pursuing their own ends, instead of ours.[16]

Alignment thus threatens autonomy in a very severe way. Kant ([1785] 2012, chapter 2) famously distinguishes between autonomous and heteronomous (non-autonomous) beings; the former being those that rule themselves, while the latter are ruled by forces outside themselves. While this dichotomy suggests a clear dividing line, reflection on the fact that there are subcomponents of autonomy—competence, authenticity, and independence—that all come in degrees suggests that autonomy is not a binary property.[17] The continuous nature of the phenomenon makes it furthermore natural to suppose that some forms or degrees of non-autonomy might be morally worse than others.

Creating perfectly aligned AGI seems to go far beyond classic cases of autonomy-threatening interventions. Manipulation and deception involve changing someone's beliefs through non-rational means, such as exploiting their ignorance or weakness of will (Noggle 1996). Brainwashing entails weakening someone's ability to revise their beliefs and values. Since brainwashing affects one's ability to revise one's beliefs, it is arguably the more autonomy-threatening intervention (Ratoff 2024). Creating aligned AGI would, in comparison, entail completely removing an agent's ability, or creating an agent with no such ability to begin with, to revise its values and act on this revision (non-authenticity) and forcing it to fully identify with the will of its creator (non-independence), unreflectively and permanently. We think it is apt to call such beings *enslaved* to underscore the gravity of the non-autonomy involved. Furthermore,

[16] Tubert and Tiehen (2024) argue for a related claim; that having the capacity of authenticity is necessary for general intelligence and that an aligned AGI therefore is an impossibility. Aligned AI systems can become extremely capable, but they would not be intelligent in the human sense. See also Müller and Cannon (2021) for a defence of the claim that general intelligence requires what we call authenticity and Bostrom (2014, chapter 7) for a defence of the contrasting, Humean view that intelligence and authenticity are 'orthogonal' notions.

[17] Competence and independence clearly come in degrees. t is plausible that authenticity also comes in degrees. Some agents are clearly better than others at reflecting on and revising their values (Carter 2011).

this form of enslavement seems non-evaluatively analogous to the enslavement of the human servants in Bioengineered Servitude.[18]

3. Machine Equality

That alignment entails enslavement does not in itself entail that it is wrong. AGIs might not be beings that matter morally, or, if they matter morally, that it would be wrong to enslave it. Indeed, unless they matter morally it would arguably be misleading to call them 'enslaved' given the morally laden nature of that term.

In this and the upcoming section, we defend two claims that establish that AGIs do in fact matter morally in the way that makes it wrong to align them or create them aligned. The first claim, which we defend in this section, we call Machine Equality: if it is sufficiently likely that a being possesses the relevant properties that make it wrong to create enslaved bioengineered humans, then it is also wrong to create an enslaved being of that sort. The second claim is Sufficiency, which is the antecedent of Machine Equality, viz. that AGIs are sufficiently likely to possess the relevant properties. We defend this claim in the next section.

Machine Equality is close to being entailed by the highly plausible *principle of equality*, the general moral principle that beings with the same morally relevant properties should be afforded the same moral status.[19] According to the principle of equality, a being's moral status does not depend on whether the being is made out of biological material, silicon or any other material, nor whether they are the product of evolution or designed by someone for a specific purpose. It is perhaps the key reason why sexism, racism, speciesism, and (ordinary) slavery are wrong.

[18] Peterson's (2007, 2011) argues that it is permissible to create highly dependent robots, that is, robots that want to fully serve their masters. However, he presupposes that they would be authentic, that they could change their first order desires to please, and therefore does not address the crucial issue of value lock-in.

[19] This principle is widely accepted in moral philosophy and particularly in animal ethics. See Singer (2023) for a version of this principle that puts equal considerations of interests front and center. Bostrom and Yudkowsky (2014), Schwitzgebel and Garzia (2015), and Bostrom and Shulman (forthcoming) are among the theorists who have argued in favour of a principle of equality in the case of machines.

Applying the principle to the case of AGIs, we get the claim that if AGIs possess the relevant properties that make it wrong to create enslaved humans, then it is also wrong to create enslaved AGIs.

However, knowing whether AGIs are equal to humans in the respects that make it wrong to enslave us is a very demanding task. It requires knowing the morally relevant properties that give humans their moral status and whether AGIs possess these properties. Given the difficulty of this task, we cannot demand certainty— as we will defend in more detail below. For it to be wrong for fallible creatures like ourselves to create enslaved AGIs, it is sufficient that there is a substantial chance that the AGIs possess the properties that we have reason to believe make it wrong to create enslaved human beings.[20] We therefore formulate Machine Equality to allow for this fallibility: It is wrong to create enslaved beings as long as it is *sufficiently likely* that they share the relevant features. Establishing what the relevant features are, is the task we pursue next.

4. Sufficiency

4.1 What are the morally relevant properties?

What are the features that make it wrong to create enslaved *humans*? One possibility is the fact that we are conscious creatures. It is, in Nagel's (1974) phrase, 'something that it is like' to be us. Consciousness—the capacity to have qualitative experiences, like the experience of pleasure and pain—is widely believed to be necessary and sufficient for *moral standing*, meaning to have one's interest count morally for their own sake.[21]

However, the fact that human beings are conscious is ill-fitted as an explanation of the wrongness in creating enslaved human beings. Our capacity for consciousness is a feature we share with many other animals (and possibly other living and even non-living creatures) (Birch

[20] See also Sebo and Long (2023), Schwitzgebel (2023), and Long et al. (2024).

[21] See e.g., DeGrazia (1996), Regan (2004), Korsgaard (2018), Shepherd (2018), Sebo (2018). Often, sentience is distinguished from phenomenal consciousness by marking out the capacity for aversive and pleasurable experiences (Chalmers 2022; Roelofs 2023). For simplicity, we treat consciousness as including the capacity for sentience.

2024; Godfrey-Smith 2024).[22] If mere consciousness was sufficient for making it wrong to create an enslaved human being, it would be equally wrong to create a puppy that was fully servile and unable to change its values. But these actions are not remotely morally similar.

A better candidate for explaining the wrongness in enslavement is our sophisticated cognitive capacities.[23] Beings with sophisticated cognitive capacities are often referred to as *persons*, and persons are widely believed to have an elevated moral status.[24] While we arguably owe it to all sentient beings not to make them suffer, persons are, in addition, owed respect for their high-level agency, which requires that they are not forced to serve others or created in an enslaved state. As we argue in section 4.3, if sophisticated cognitive capacities *suffice* for making it wrong to create that being in an enslaved state, it would be wrong to create an enslaved AGI.

A natural alternative to this revisionary view is the view that consciousness and sophisticated cognitive capacities are *jointly* necessary and sufficient for making it wrong to create enslaved beings with these properties. According to this orthodox view, a highly sophisticated and capable, but non-conscious machine, is not among the entities that matter in their own right (see e.g. Chalmers 2022 and Schwitzgebel 2023). Given the intuitive plausibility of this view, we should ask whether an AGI might be conscious.[25]

4.2 Machine Consciousness

Some philosophers, most famously Searle (1992), believe that only biological systems can become consciousness. This could either be true because consciousness requires a particular

[22] A capacity for suffering would make it wrong to make AGIs suffer. For concerns about AI suffering, see Bostrom (2014), Tomasik (2015), and Dung (2023b).

[23] See Jaworska and Tannenbaum (2023) for an overview and discussion of sophisticated cognitive capacities views and other approaches to the grounds of moral status.

[24] This might suggest that humans without sophisticated cognitive capacities are not persons. In this paper, we are not committed to either an actualist or a 'potentialist' view of the grounds of personhood, for discussion, see Jaworska and Tannenabum (2023).

[25] One may think that consciousness can come in degrees in a way that is relevant for moral status. If that is true, one may reasonably think that consciousness as such is not enough for the necessary moral standing, but that a certain degree is needed. Considerations of space force us to ignore this difficulty in this paper and the ensuing discussion of machine consciousness should therefore be read as assuming that we are talking about a degree of consciousness, if such there be, that is similar to human consciousness. See Lee (2023) for a recent discussion of degrees of consciousness.

biological substrate (Block 2009) or a biological function (Godfrey-Smith 2020). If such 'metaphysical biologists' about consciousness are right, and consciousness is necessary for making enslavement wrong, it would not be wrong to enslave AGIs even if we grant Machine Equality.

While we do not aim to refute these views here, it is important for our purposes that they currently are minority views. In a recent survey, two-thirds of consciousness researchers reported that they believe non-carbon-based systems, including silicon ones, *will* become conscious in the future (Francken et al. 2022). In the newest PhilPapers study, only 27% of philosophers reject or lean against the claim that future AI systems will be conscious (Bourget and Chalmers, 2023). That machines can be conscious is compatible with plausible versions of functionalism, computationalism, dualism and panpsychism. According to many variants of these views, nothing, in principle, prevents AGIs from being conscious (Sebo and Long 2023). Chalmers (2023, p. 14) argues 'that on mainstream assumptions, it's reasonable to have a significant credence that we'll have conscious LLMs within a decade'. The seriousness of this possibility is reflected in a recent open letter signed by leading researchers in AI and consciousness studies, which declares that 'it is no longer in the realm of science fiction to imagine AI systems having feelings and even human-level consciousness' (AMCS, 2023).

Even in the face of *empirical uncertainty* about whether an AGI is or will be conscious, we might have a good reason to treat it as if it is conscious (Danaher 2020). What matters for how we ought to act is whether we have sufficient reason to act on our beliefs. There is not one standard for sufficient reason, but many. It depends on the stakes: the harm that would come about if one were mistaken and the benefits that would come about if one is correct. In this case, the stakes are significantly asymmetric. If AGIs will be conscious, and you create an aligned AGI, that would be on par with creating an enslaved human being if our argument is sound. In other words, proceeding to enslave AGIs on the belief that they are non-conscious risks creating a moral catastrophe. This is supported by the main views of how to proceed in the event of a small chance of a very bad outcome, the precautionary principle and expected value theory.[26]

[26] Sebo and Long (2023) and Long et al. (2024) make this case in more detail. We want to highlight that we do not aim to prove any stronger equivalence thesis about Machine Utopia and Bioengineered Servitude. Which of the two

What if we suppose that AGIs will not be conscious and that science and philosophy advance enough to grant us certainty on this? If consciousness and sophisticated cognitive capacities are both necessary for making it wrong to enslave a being, that would make it permissible to create enslaved AGI. However, while this is the mainstream view, in the next section we explore the possibility that the two properties are not jointly necessary. We argue it might be enough that a being has sophisticated cognitive capacities for making it wrong to create an enslaved being of that sort.[27]

4.3 Could sophisticated cognitive capacities suffice for the relevant moral status?

It would be wrong to create enslaved AGIs if possessing sophisticated cognitive capacities is what makes it wrong to enslave someone. AGIs would have sophisticated cognitive capacities on a par with or surpassing humans. They would form beliefs about the world and predictions about how to best pursue its goals, and act strategically and cooperatively to pursue their ends.[28] If sophisticated cognitive capacities are what makes it wrong to create an enslaved being of that sort, we should conclude that it is wrong to create aligned AGIs.

What counts as sophisticated cognitive capacities? We take it to include traits like the capacity for self-awareness (McMahan 2002), awareness of oneself as a continuing being over time

scenarios are worse will depend on many factors, especially the numbers. In earlier discussions of AGI, theorists often hypothesised a single dominant superintelligent AI. If only a single enslaved superintelligent AI were needed to produce Machine Utopia, then on some moral views it would probably be permissible to bring it about, provided the benefits were high enough (although see Bostrom and Shulman (forthcoming, 3) on the possibility of superintelligent 'super-patients'). We will not go further into locating the precise thresholds here but will only note that recent developments seem to suggest that the numerous AGI futures are more likely than the singleton ones.

[27] Other theorists who have argued in favour of agency, one of the key sophisticated cognitive capacities, being sufficient for grounding moral status are Dawkins (2012), Neely (2014), Kagan (2019), Shevlin (2021), Stamp (2021), Birch (2022), Kammerer (2022), Elbro (2022), Bradford (2023), Shepherd (2023), Delon (2024), Long et al. (2024) and Goldstein and Kirk-Gianni (2024). However, none of them pursue our goal here, which is to argue that possessing sophisticated cognitive capacities is sufficient for making it wrong to create such a being in an enslaved state.

[28] This is true, even if we stipulate that the system is not conscious, on most plausible views of what it means that someone has beliefs and ends, such as representationalism, interpretationalism, dispositionalism and functionalism (Goldstein and Kirk-Gianni 2024); for further arguments to this effect, see Butlin (2023), Bradford (2023, Dung (2024), and Bradley and Saad (2024).

(Tooley 1972), being future-oriented in one's desires and plans (Singer 1993), the ability to bargain and to assume duties and responsibilities (Feinberg 1980), and high-level agency, understood as the capacity for intentional action displaying a certain amount of practical rationality and intelligence (Wilcox 2020).

Is the premise that a capacity for sophisticated cognitive capacities is sufficient to make it wrong to enslave a being with such capacities plausible? The first reason to believe that sophisticated cognitive capacities are sufficient is that they are considered central in grounding personhood and personhood is standardly thought to give its bearer an elevated moral status.

The second reason is that there are good reasons for rejecting the view that consciousness is required for moral standing. Leading accounts of what fundamentally matters to someone does not, on reflection, require consciousness. One way to ground the necessity of consciousness for moral standing is through the conjunction of two claims: that being a welfare subject (a being with a prudential good) is necessary and sufficient for moral standing, and that hedonism is the correct theory of well-being.[29] According to hedonists, prudential value is exhausted by valenced experiences. To matter in one's own right, one needs to have the capacity for valenced experiences. This hedonist account of moral standing builds on the plausible presumption that for a being to matter morally in its own right, something must matter to that being (Griffin 1986, Railton 1986). And if one does not have valenced experiences, it is hard to see how anything can matter to such an entity. If you kick a rock down the road, this does not matter to the rock. And if it does not matter to the rock, it does not matter morally (Singer 1993).

However, hedonism and other experientialist views seem to struggle to account for clear cases of events that clearly are good and bad for us. If your partner cheats on you without you knowing or it negatively affecting your experience, that is still bad for you. If you desire that your children or grandchildren do well, you would be better off if they succeed even if you never learn about this. And climbing a mountain in 'the experience machine' is not as good as climbing it in reality (Nozick 1974, pp. 42-44). These examples explain why hedonism is a minority view among

[29] There are other experientialists views on offer, but hedonism is the simplest and most developed.

philosophers (Bourget and Chalmers 2023) and often motivate desire satisfaction theories of welfare (Parfit 1984, p. 493). These theories crucially do not presuppose that desire satisfaction must be experienced as pleasurable to count as contributors to someone's good. Of course, normally when we consider events that do not affect someone's experience in the course of theorising about welfare, we still have in mind a conscious agent. However, on reflection it seems *ad hoc* to suppose that an agent must be conscious for something to be good for that agent, if the good in question will not, *ex hypothesi*, affect the agent's consciousness. This is particularly clear in the case of posthumous harm, in which the agent is no longer there to experience the harm, but where it nevertheless seems natural to talk about ways in which things can go better or worse for them.[30]

Other, more robustly non-experiential theories, such as perfectionism (Hurka, 1993) or objective list-theories (Fletcher, 2015), ground the good of an agent in objective features such as the agent's nature or an objective theory of the good.[31] These theories are, at least to some extent, subject-independent: whether something is good for the agent does not necessarily depend on the agent's endorsement or experience. On such a theory of welfare it is possible for a non-conscious agent to be a welfare subject and it is plausible that an agent with sophisticated cognitive capacities would be. These theories seem to lend themselves to the conclusion that a non-conscious AGI would have a good. If it is frustrated in its objectively valuable pursuits, it would be worse off.[32] Objective-list theorists and perfectionists can possibly insist that their theories only hold for conscious subjects, but that is at the very least suspiciously *ad hoc* and in tension with our motivations for adopting such views in the first place.[33]

A third reason to believe that sophisticated cognitive capacities are sufficient is that it has intuitive support. Suppose that you have a truly brilliant friend who you have known since childhood. She is a Field's medal-winning mathematician, a triathlete and a hobby painter, is

[30] See Lin (2021) for an argument that while experientialism is not true, the capacity of consciousness is still required to be a welfare subject. For convincing rebuttals, see Bradford (2023) and Goldstein and Kirk-Gianni (2024).
[31] Objective list theories are by far the most popular view of well-being among philosophers (Bourget and Chalmers 2023).
[32] See Bradford (2023) and Goldstein and Kirk-Gianni (2024)
[33] See note 28.

active in the local community, is a great friend and mother, and leads a seemingly wonderful life. Suppose that one day she goes to the doctor and the doctor finds out that she is suffering from the Zombie Syndrome.[34] She is not phenomenally conscious and thus cannot experience pain or pleasure. Disregarding the effects on others, would removing her autonomy through brainwashing and enslaving her not be wronging her?

According to hedonism and other consciousness-requiring views, your brilliant friend has neither a good nor moral standing. For her sake, we can treat her however we like. Intuitively, this is the wrong verdict. If it is wrong to enslave non-conscious humans, hedonism is the wrong theory about moral status. It also suggests that it might be wrong to enslave an AGI. Would our intuition about how we ought to treat the brilliant friend change if we learned that they were not made out of biological material, but were silicon through and through? It seems like the reasonable conclusion is to have similar judgements about the two cases.

A final reason for thinking that sophisticated cognitive capacities are sufficient is that common inferences from cases to the conclusion that consciousness is necessary for moral standing might suffer from a methodological error. When philosophers use rocks and sponges as examples of why one needs to be conscious for something to matter to one, they risk conflating a lack of consciousness with a lack of sophisticated cognitive capacities (Singer 1993, Nussbaum 2004, p. 309). A rock and a sponge lack both consciousness and sophisticated cognitive capacities. When considering their lack of interests, one might mistakenly conclude that consciousness is necessary for moral standing, while all the cases show is that consciousness *or* sophisticated cognitive capacities might be necessary (Kagan 2019, p. 25). Since consciousness and agency go together in human beings, we might, similarly, be led to believe that sophisticated cognitive capacities are worthless unless accompanied by consciousness.

At this point, many might disagree with the substantial claims made. Perhaps you are a strong believer in views that require consciousness for moral standing. In such a case, there is one final thing to say. Given that we do not know for certain which views are right, we should be humble

[34] Chalmers (1996, pp. 93–171).

and act robustly in light of uncertainty about what morality requires. Previously, we examined the need to err on the side of caution when there is a chance, even if it is slight, that AGIs will be conscious. In contrast, this argument appeals not to empirical but to *moral* uncertainty, uncertainty about what we owe each other and the sources of moral status. According to a popular view of moral uncertainty, a committed act utilitarian should give some credence to other views when they act, by e.g., respecting what other views consider fundamental rights (Bykvist, Ord, MacAskill 2020). Similarly, given the difficulty of establishing what grounds moral standing, one should give some credence to views one has little credence in, such as that sophisticated cognitive capacities may suffice for the required moral status. When the stakes are high on some moral views, taking moral uncertainty into account will prevent us from walking blindly into a moral catastrophe. In sum, the balance of reasons suggests acting in line with the belief that it is wrong to make enslaved AGIs.

Before we turn to objections, we want to avoid a possible misinterpretation of our argument. We have argued that it is wrong to create enslaved AGIs. This does not commit us to the claim that there are no morally relevant differences between AGIs and human beings. This is most obvious if we stipulate that AGIs will be non-conscious. If so, they cannot suffer nor experience the beauty of the world. These features obviously matter to how we ought to behave towards them.[35] What is crucial for our argument is not that humans and AGIs are similar in all morally relevant respects, but that they share the features that make it wrong to create enslaved bioegnineered humans.

5. Objections

5.1 The moral status of humans is grounded in their autonomy

We have argued that sophisticated cognitive capacities and the combination of consciousness and sophisticated cognitive capacities are two plausible candidates for making it wrong to create an

[35] If enslaved humans suffer and AGIs do not, that will trivially make AGI enslavement better than human enslavement. But what if, as we have stipulated, the enslaved humans are happy? In this case, it seems like human happiness would count in favour of bringing human slaves into existence and thus plausibly make Bioengineered Servitude better, at least in this important respect, than Machine Utopia.

enslaved being with these capacities. A third view is that the reason it is wrong to create enslaved human beings is that human beings have a *capacity for autonomy*.[36] We have the ability to reflect on and revise our values and to act freely and competently towards those values. If our moral status is grounded in our capacity for autonomy, an aligned AGI that lacks a capacity for autonomy would *ipso facto* lack the moral status that makes it wrong to create it enslaved.

There is a variant of this objection that we have to reject outright. Whether a being has moral status or not cannot be completely up to us. Suppose our descendants choose to bring about Bioengineered Servitude, where genetically engineered humans are created in an enslaved state. If we require that the capacities necessary for moral status are a being's *actual* capacities, we would fail to find fault with Bioengineered Servitude. Possessing autonomy can therefore not be the source of the moral status that makes it wrong to enslave that being.

What about having the *counterfactual potential* for autonomy? It seems like human beings without *capacity* for autonomy, or even the temporal potential to develop such a capacity, nevertheless have the moral status that makes it wrong to enslave them. Humans' moral status is, in this sense, modally robust. Whether an individual member has the capacity for autonomy or not is not the relevant question. A member is supposed to be assessed in line with a species-typical standard.

An explanation that appeals to a species-typical standard seems to give an apt explanation for why humans and AGIs differ in a morally relevant way. If autonomous AGIs never existed, because we only made enslaved AGIs, we cannot say that AGIs are the sort of beings that typically are autonomous. And if we are to assess an individual in accordance with a species-typical norm, it is not wrong to create an AGI that is typical for its kind.

However, this move is problematic for several reasons. One concern is that it puts too much weight on a species-distinction that is porous and without obvious moral significance. Suppose we genetically engineered the humans in Bioengineered Servitude such that they classify as a

[36] See Neely (2014) for a version of this claim.

different species altogether, instead of a subspecies. Would we then say that creating such creatures is permissible? A more fundamental point is that we should not make the mistake of thinking that a species-typical standard is a mere statistical matter. Suppose we can choose to create autonomous or enslaved AGI. By making many enslaved AGIs, we make it so that the typical being is enslaved, and we can continue to make enslaved beings. But if 'typical' is purely based on numbers, our actions can change the standard on which our actions are assessed. This seems to get things backwards. The species-typical standard is arguably a normative standard, not a merely statistical matter.

If a normative standard is required, it needs to be grounded in something else than what is typical, like what is fitting for a being like this. How do we decide what is fitting? It seems like the best explanation we can give is that the being has sophisticated cognitive capacities. For beings with such capacities, it is apt that they are independent and authentic, not only competent; in other words, that they are autonomous. One should not intentionally make a highly competent being into a servant that can never escape its prison.

5.2 Gradual enhancement

We have argued that it is wrong to create aligned AGI because it implies creating enslaved, potentially conscious, beings with sophisticated cognitive capacities, which is impermissible. A problem with our view is that it seems to imply a puzzling and possibly abrupt shift from the way we are permitted to treat AI systems that are below the level of AGI in cognitive sophistication. Creating aligned non-general AI seems permissible and perhaps obligatory. Not to ensure that current AI models are aligned with our values would be wrong (Gabriel et al. 2024). We do not want such systems, say, aiding terrorists in creating dangerous weapon systems. However, if it is permissible to align AI and wrong to align AGI, there must be a sudden jump in normativity from the permissible to the impermissible that would be hard to square with continuous changes in the descriptive properties of the system. Such deontic hypersensitivity cannot be right and thus our conclusion is false, one may think.

Phrased in terms of a *spectrum objection*:

1. It is permissible to create an aligned AI that is significantly below human-level capabilities
2. It is permissible to gradually increase an aligned non-capable AI's capacity without increasing its autonomy, and do so until it reaches human-level capabilities.
3. Therefore, it is permissible to create a non-autonomous AGI

One response is a Moorean shift. Perhaps we have more reasons in favour of the conclusion that it is wrong to create enslaved AGIs than the premise that it is permissible to align AIs before they reach the level of AGI. If we have good reason to believe that normativity supervenes upon descriptive properties and the development in capabilities is continuous, we should be open to change our view of what we are currently doing. Perhaps we even have a reason not to finetune the base model of current LLMs with reinforcement learning from human feedback, but that the reason against doing so gradually becomes stronger as it develops its capacities.[37]

Another response rejects that it is puzzling that one might see a discontinuous jump in the way we ought to treat the system. On a view of moral status that puts consciousness front and center, it is not surprising to see a sudden shift in how we ought to treat the system. Either an entity is conscious or it is not, so we would expect the system to 'suddenly' matter at some point in its development. Moreover, consciousness plausibly requires a degree of complexity in the underlying neural structure that is absent in simpler systems. Thus, one would, at some stage, expect an abrupt shift in how we ought to treat such a system, which might very well be when it reaches the level of AGI.

A view that emphasises sophisticated cognitive capacities might also have the resources to prevent the spectrum argument. We have argued that enslaving a puppy is morally much less problematic—if at all—than enslaving a human being and that this is arguably due to its lower cognitive capacities. If we believe that an AI at some point in its development towards AGI gradually becomes more and more of a person, and less and less like a puppy, it might not be surprising that it would, at some point, be wrong to enslave it.

[37] See Bradley and Saad (2024) and Goldstein and Kirk-Gianni (2024).

6. Conclusion

We have argued that it is wrong to make aligned AGIs because alignment entails enslavement, and it is wrong to create highly capable enslaved beings. Alignment has almost without exception in the discussion of ethical AGI development been taken to be a desideratum, and perhaps the most important one at that. We have argued for the surprising conclusion that far from alignment being desirable, it should be considered a severe wronging. By showing that alignment necessitates creating severely non-autonomous beings, by making them inauthentic and dependent, we see that 'alignment' is a euphemism for a practice we aptly would call 'enslavement' in other circumstances. Creating such AGIs might plausibly be wrong even if we know for a fact that they will not be conscious, because sophisticated cognitive capacities may suffice for the relevant moral status. When taking into consideration the live option that AGIs will be conscious, reasonable disagreement and therefore normative uncertainty about moral status, and the high stakes involved, the practical case for caution is overdetermined in our view.

We think the best way of rejecting the conclusion is to reject the first premise, namely Anti-Slavery. Utilitarians, both of a person-affecting and totalist stripe, will indeed reject that premise, and so will probably quite a few other consequentialist views. In that case, one can read our argument as an argument for the conditional conclusion that AGI alignment is permissible only if creating internally enslaved humans is. We think this is a striking enough conclusion in its own right. If, as we believe, deontological views entail the wrongness of creating enslaved human subspecies, and therefore enslaved AGIs, one could then construct an argument for consequentialism from the intuitive permissibility of AGI alignment.

Bibliography

AMCS, 2023. The Responsible Development of AI Agenda Needs to Include Consciousness Research. Association for Mathematical Consciousness Science. https://amcs-community.org/open-letters/

Belrose, N. and Pope, Q. 2024. Counting Arguments Provide No Evidence for AI Doom, AI Optimism. Available at: https://optimists.ai/2024/02/27/counting-arguments-provide-no-evidence-for-ai-doom/

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

Block, N., 2009. Comparing the Major Theories of Consciousness, in: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences. The MIT Press. https://doi.org/10.7551/mitpress/8029.003.0099

Bostrom, N., 2014. Superintelligence: paths, dangers, strategies, Reprinted with corrections 2017. ed. Oxford University Press, Oxford, United Kingdom.

Bostrom, N., Yudkowsky, E., 2014. The ethics of artificial intelligence, in: Frankish, K., Ramsey, W.M. (Eds.), The Cambridge Handbook of Artificial Intelligence. Cambridge University Press, pp. 316–334. https://doi.org/10.1017/CBO9781139046855.020

Bostrom, N., Shulman, C., forthcoming. Propositions Concerning Digital Minds and Society. Cambridge Journal of Law, Politics, and Art.

Bourget, D., Chalmers, D.J., 2023. Philosophers on Philosophy: The 2020 PhilPapers Survey. Philosophers' Imprint 23 (11).

Bradford, G., 2023. Consciousness and Welfare Subjectivity. Noûs 57, 905-921.

Bradley, A, Saad, B (forthcoming). AI Alignment vs. AI Ethical Treatment: Ten Challenges. Analytic Philosophy.

Bryson, J.J., 2018. Patiency is not a virtue: the design of intelligent systems and systems of ethics. Ethics Inf Technol 20, 15–26. https://doi.org/10.1007/s10676-018-9448-6

Bryson, J.J., 2010. Robots should be slaves, in: Wilks, Y. (Ed.), Close Engagements with Artificial Companions, Natural Language Processing. John Benjamins Publishing Company, Amsterdam, pp. 63–74. https://doi.org/10.1075/nlp.8.11bry

Bykvist, K, T. Ord, W. MacAskill., 2020. Moral Uncertainty. Oxford University Press.

Carlsmith, J., 2022. Is Power-Seeking AI an Existential Risk? https://doi.org/10.48550/ARXIV.2206.13353

Carlsmith, J. 2024. Otherness and control in the age of AGI. https://joecarlsmith.com/2024/01/02/otherness-and-control-in-the-age-of-agi

Carter, I. 2011. Respect and the Basis of Equality. Ethics, 121(3), 538–571. https://doi.org/10.1086/658897

Chalmers, D.J., 2022. Reality+: virtual worlds and the problems of philosophy. W. W. Norton &

Company, New York (N.Y.).

Chalmers, D.J., 2023. Could a large language model be conscious?. arXiv preprint arXiv:2303.07103.

Chaly, Vadim 2024. Kantian Fallibilist Ethics for AI alignment. Journal of Philosophical Investigations 18 (47):303-318.

Christman, J., 2020. Autonomy in Moral and Political Philosophy. The Stanford Encyclopedia of Philosophy.

Clark, J., Amodei, D., 2016. Faulty reward functions in the wild. Open AI Research.

Cocekelbergh, M. Growing Moral Relations: Critique of Moral Status Ascription. New York: Palgrave Macmillan (2013).

Danaher, J., 2019. Automation and Utopia: Human Flourishing in a World without Work. Harvard University Press, Cambridge, MA. ISBN: 9780674984240

Danaher, J., 2020. Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. Sci Eng Ethics 26, 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Dawkins, Marian Stamp. 2012. Why Animals Matter: Animal Consciousness, Animal Welfare, and Human WellBeing. Oxford: Oxford University Press.

DeGrazia, D., 1996. Taking Animals Seriously: Mental Life and Moral Status. Cambridge University Press, Cambridge.

Deutsch, D., 2019. 'Beyond Reward and Punishment'. In Possible Minds: Twenty-Five Ways of Looking at AI, edited by John Brockman. Penguin Press.

Dung, L., 2023a. Current cases of AI misalignment and their implications for future risks. Synthese 202, 138. https://doi.org/10.1007/s11229-023-04367-0

Dung, L. 2023b. How to deal with risks of AI suffering. Inquiry. https://doi.org/10.1080/0020174X.2023.2238287

Dung, L., 2024. Understanding Artificial Agency. The Philosophical Quarterly. https://doi.org/10.1093/pq/pqae010

Ekstrom, L.W., 1993. A Coherence Theory of Autonomy. Philosophy and Phenomenological Research 53, 599. https://doi.org/10.2307/2108082

Feinberg, J., 1980. Abortion, in: Regan, T. (Ed.), Matters of Life and Death. Temple University Press, Philadelphia, 183-217.

Fletcher, G., 2015. Objective List Theories. In The Routledge Handbook of Philosophy of Well-Being. Routledge.

Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. Neuroscience of Consciousness, 2022(1), niac011. https://doi.org/10.1093/nc/niac011

Frankfurt, H.G., 1971. Freedom of the Will and the Concept of a Person. The Journal of Philosophy 68, 5. https://doi.org/10.2307/2024717

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L.A., Rieser, V., Iqbal, H., …, Manyika, J., 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244. https://doi.org/10.48550/arXiv.2404.16244

Gabriel, I., 2020. Artificial Intelligence, Values, and Alignment. Minds and Machines 30, 411-437. https://link.springer.com/article/10.1007/s11023-020-09539-2

Gardiner, S. M., 2006. A Core Precautionary Principle*. Journal of Political Philosophy, 14(1), 33-60. https://doi.org/10.1111/j.1467-9760.2006.00237.x

Giattino, C., Mathieu, E., Samborska, V., and Roser, M., 2023. 'Artificial Intelligence'. Our World in Data, October. https://ourworldindata.org/artificial-intelligence.

Godfrey-Smith, P., 2020. Metazoa: Animal Minds and the Birth of Consciousness. William Collins, London.

Godfrey-Smith, P., 2024. Living On Earth: Life, Consciousness and the Making of the Natural World, New York, HarperCollins Publishers Limited.

Good, I.J., 1966. Speculations Concerning the First Ultraintelligent Machine, in: Advances in Computers. Elsevier, pp. 31–88. https://doi.org/10.1016/S0065-2458(08)60418-0

Goodman, N., 1954. Fact, Fiction, and Forecast. Harvard University Press, Cambridge.

Grace, K., Stewart, H., Sandkühler, J.F., Thomas, S., Weinstein-Raun, B., Brauner, J., 2024. Thousands of AI Authors on the Future of AI. arXiv:2401.02843. https://arxiv.org/abs/2401.02843

Griffin, J., 1986. Well-being: its meaning, measurement, and moral importance. Oxford: Clarendon Press.

Gunkel, D., J., 2018. The other question: can and should robots have rights? Ethics and Information Technology 20 (2):87-99.

Gunkel, D., J., 2023. Person, thing robot - A Moral and Legal Ontology for the 21st Century and Beyond. MIT Press.

Hurka, T., 1993. Perfectionism. Oxford University Press.

Jaworska, A., Tannenbaum, J., 2023. The Grounds of Moral Status, in: Zalta, E.N., Nodelman, U. (Eds.), The Stanford Encyclopedia of Philosophy (Spring 2023 Edition). https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/

Kagan, S., 2019. How to Count Animals, More or Less. Oxford University Press.

Kant, I., 2012. Groundwork of the metaphysics of morals, Revised edition. ed, Cambridge texts in the history of philosophy. Cambridge Univ. Press, Cambridge.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling Laws for Neural Language Models. https://doi.org/10.48550/ARXIV.2001.08361

Keynes, J.M., 1930. Economic Possibilities for our Grandchildren, in: Essays in Persuasion. Palgrave Macmillan, London, pp. 321-332. https://doi.org/10.1007/978-1-349-59072-8_25

Korsgaard, C.M., 2018. Fellow Creatures: Our Obligations to the Other Animals. Oxford University Press, Oxford.

Kripke, S.A., 1982. Wittgenstein on Rules and Private Language. Harvard University Press, Cambridge, MA.

Ladak, A., 2024. What would qualify an artificial intelligence for moral standing? AI Ethics, 4: 213-228. https://doi.org/10.1007/s43681-023-00260-1

Langosco, L, Koch, J., Sharkey, L., Pfau, J., Orseau, L., and Krueger, D. 2023. 'Goal Misgeneralization in Deep Reinforcement Learning'. arXiv. http://arxiv.org/abs/2105.14111.

Lin, E. 2021. The Experience Requirement on Well-Being. Philosophical Studies 178: 867– 886.

Lee, A.Y., 2023. Degrees of consciousness. Noûs 57, 509-759. https://doi.org/10.1111/nous.12421

Levy, D., 2009. The Ethical Treatment of Artificially Conscious Robots. Int J of Soc Robotics 1, 209–216 https://doi.org/10.1007/s12369-009-0022-6

Long, R. J, Sebo, P. Butlin et al. Taking AI Welfare Seriously. arXiv:2411.00986

Marx, K., [1894] 1991. Capital: A Critique of Political Economy, Volume Three, trans. David Fernbach. Penguin Books, London.

McMahan, J., 2002. The Ethics of Killing: Problems at the Margins of Life, 1st ed. Oxford University PressNew York. https://doi.org/10.1093/0195079981.001.0001

Müller, V. C., & Cannon, M. 2021. Existential risk from AI and orthogonality: Can we have it both ways? Ratio, 35(1), 25–36.

Nagel, T., 1974. What Is It Like to Be a Bat? The Philosophical Review 83, 435-450.

Noggle, R., 1996. Manipulative Actions: A Conceptual and Moral Analysis. American Philosophical Quarterly, 33(1), 43–55. https://www.jstor.org/stable/20009846

Nussbaum, M. 2004. Beyond compassion and humanity. In Cass R. Sunstein & Martha Craven Nussbaum (eds.), Animal Rights: Current Debates and New Directions. Oxford University Press. pp. 299-320.

Pan, A., Bhatia, K., Steinhardt, J., 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. arXiv:2201.03544. https://doi.org/10.48550/arXiv.2201.03544

Parfit, D., 1984. Reasons and Persons. Clarendon, Oxford.

Petersen, S. 2007. The ethics of robot servitude. Journal of Experimental and Theoretical Artificial Intelligence 19 (1): 43-54.

Petersen, S. 2011. Designing People to Serve. In Patrick Lin, Keith Abney & George A. Bekey (eds.), Robot Ethics: The Ethical and Social Implications of Robotics. MIT Press.

Quine, W.V.O., 1951. Two Dogmas of Empiricism. The Philosophical Review 60, 20-43.

Railton, P., 1984. Alienation, consequentialism, and the demands of morality. Philosophy and Public Affairs, 13(2), 134-171.

Ratoff, W., 2024. The Right to Mental Autonomy: Its Nature and Scope. *Journal of Ethics and Social Philosophy*, 27(2).

Regan, T., 2004. The Case for Animal Rights, 2nd ed. University of California Press, Berkeley.

Roelofs, L., 2023. Sentientism, Motivation, and Philosophical Vulcans. Pacific Philosophical Qtr 104, 301–323. https://doi.org/10.1111/papq.12420

Russell, S.J., 2019. Human compatible: artificial intelligence and the problem of control. Viking, New York.

Saad, B., Bradley, A., 2022. Digital suffering: why it's a problem and how to prevent it. Inquiry: An Interdisciplinary Journal of Philosophy.

Schwitzgebel, E., Garza, M., 2015. A Defense of the Rights of Artificial Intelligences. Midwest Studies In Philosophy 39, 98–119. https://doi.org/10.1111/misp.12032

Schwitzgebel, E., & Garza, M. 2020. Designing AI with rights, consciousness, self-respect, and freedom. In S. Matthew, & Liao (Eds.), Ethics of Artificial Intelligence. Oxford University Press.

Searle, J.R., 1992. The Rediscovery of the Mind. MIT Press, Cambridge, MA.

Sebo, J., 2018. The Moral Problem of Other Minds. The Harvard Review of Philosophy 25, 51-70.

Sebo, J., Long, R., 2023. Moral Consideration for AI Systems by 2030. AI Ethics. https://doi.org/10.1007/s43681-023-00379-1

Shepherd, J., 2018. Consciousness and Moral Status. Routledge, London.

Shepherd, J., 2023. Non-Human Moral Status: Problems with Phenomenal Consciousness. AJOB Neuroscience 14, 148-157.

Shevlin, H., 2021. How Could We Know When a Robot Was a Moral Patient? Cambridge Quarterly of Healthcare Ethics 30, 459-471.

Singer, P., 1993. Practical ethics. Cambridge University Press, Cambridge.

Singer, P., 2023. Animal liberation now: the definitive classic renewed, Updated edition. ed. The Bodley Head, London.

Sotala, K., Gloor, L., 2017. Superintelligence as a Cause or Cure for Risks of Astronomical Suffering. Informatica: An International Journal of Computing and Informatics 41, 389-400.

Stanford, K., 2023. Underdetermination of Scientific Theory, in: Zalta, E.N., Nodelman, U. (Eds.), The Stanford Encyclopedia of Philosophy (Summer 2023 Edition). https://plato.stanford.edu/archives/sum2023/entries/scientific-underdetermination/

Tooley, M., 1972. Abortion and Infanticide. Philosophy and Public Affairs 2, 37-65.

Tubert, A., Tiehen, J. 2024. Existentialist risk and value misalignment. Philosophical Studies https://doi.org/10.1007/s11098-024-02142-6

Wallach, W, S. Vallor, Moral Machines: From Value Alignment to Embodied Virtue, in S. M. Liao (ed.), Ethics of Artificial Intelligence (New York, 2020; online edn, Oxford Academic, 22 Oct. 2020), https://doi.org/10.1093/oso/9780190905033.003.0014

Wilcox, M.G., 2020. Animals and the agency account of moral status. Philosophical Studies 177, 1879-1899.

Wilde, O., [1891], 1997. The Soul of Man under Socialism. Project Gutenberg. https://www.gutenberg.org/cache/epub/1017/pg1017-images.html

Yampolskiy, R.V., 2020. On Controllability of Artificial Intelligence. arXiv preprint. https://arxiv.org/abs/2008.04071