Against Willing Servitude: Autonomy in the Ethics of Advanced Artificial Intelligence

Adam Bales (Global Priorities Institute, University of Oxford)

Global Priorities Institute | October 2024

GPI Working Paper No . 23-2024

Please cite this working paper as: Bales, A. Against Willing Servitude: Autonomy in the Ethics of Advanced Artificial Intelligence. *Global Priorities Institute Working Paper Series, No. 23-2024.* Available at https://globalprioritiesinstitute.org/against-willing-servitude-autonomy-in-the-ethics-of-advanced-artificial-intelligence-adam-bales

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Against Willing Servitude

Autonomy in the Ethics of Advanced Artificial Intelligence

Adam Bales

Abstract

Some people believe that advanced artificial intelligence systems (AIs) might, in the future, come to have moral status. Further, humans might be tempted to design such AIs that they serve us, carrying out tasks that make our lives better. This raises the question of whether designing AIs with moral status to be willing servants would problematically violate their autonomy. In this paper, I argue that it would in fact do so.

1 Introduction

"A love of nature keeps no factories busy," proclaims a character in Aldous Huxley's *Brave New World*. For this reason, babies from the Delta caste are shown flowers and then electrocuted, to condition them to abhor nature. After all, the Deltas are a servile caste, their lives dedicated to meeting the needs of Alphas and Betas. What use a servant who wastes time on flippancies like

flowers? More generally, the Deltas are selectively bred and subjected to conditioning and propaganda to shape them into ideal servants.

This isn't a clean thought experiment that isolates a single moral consideration. Still, one wrong here plausibly results from the Deltas being created and moulded as servants. The fact that the Deltas are treated as mere tools for fulfilling others' desires seems to impair their autonomy.

If so, concerns arise for human treatment of artificial intelligence systems (AIs). After all, humans will plausibly shape AIs' desires as we would shape a tool, selecting these so they serve human interests. Presumably this raises no moral concern for now, when AIs lack moral status and arguably lack desires too. But perhaps AIs will one day matter morally and perhaps they'll desire. It might be worried that, at that point, we'll impair the autonomy of AIs if we shape their desires in order to make them willing servants.

In this paper, I'll explore the ethics of AI servitude, arguing that the creation of willing AI servants would indeed impair these systems' autonomy. This provides *pro tanto* moral reasons against creating such AIs. All else being equal, we ought not create willing AI servants.[1]

2 Moral Status and Willing Servitude

I'll focus on AIs that both: (a) have moral status; and (b) are willing servants.

To have *moral status* is to matter morally, to some extent, for one's own sake (this is sometimes called *moral standing*; see Kagan, 2019, ch. 1). Humans have moral status as do many non-human animals. Other objects lack such status or it's unclear whether they have it. Consider rocks, bacteria, stag-headed oak trees, or active human neurons in a petri dish.

Consider current AIs. It strikes many as unlikely that these AIs matter morally.[2] Yet it also strikes many as possible that, in the future, some AIs might come to have moral status (Shulman & Bostrom, 2021, p. 306; DeGrazia, 2022).[3]

One reason to think this possibility open is that phenomenal consciousness is sometimes thought to suffice for moral status (Chalmers, 2022, pp. 158–160; Lee, Manuscript). And

---

[1] This move from *pro tanto* reasons to what we ought to do, all else equal, is justified only because our reasons to not impair autonomy are requiring reasons (see Little & Macnamara, 2020).

[2] Though some think they might (cf. Goldstein & Kirk-Giannini, manuscript).

[3] Moosavi (2023) argues there's no reason to think future AIs will have moral status. However, she sets aside the type of AIs with the strongest claim to moral status: AIs with, "the full range of human mental capacities, including phenomenal consciousness" (p. 13). I doubt we can confidently dismiss the possibility of such systems over long timeframes (as Moosavi acknowledges). In any case, Moosavi's view isn't settled consensus, and her argument relies on disputable assumptions about moral status.

prominent theories take consciousness to arise from certain cognitive or information-processing capacities.[4] Given that future AIs could, for all we know, have these capacities, we should take seriously the possibility that they might be conscious and hence have moral status.[5] Along similar lines, sentience (by which I mean, valenced consciousness) is often thought to suffice for moral status (Singer, 2015, p. 23). And if AIs might be conscious then it's reasonable to take seriously the possibility that they could also be sentient and hence have moral status.

More theory-agnostically, there's uncertainty and disagreement regarding consciousness, sentience, moral status, and the theoretical and practical limits of AI. Further, AIs could come to possess, in an increasingly sophisticated manner, many of the markers associated with paradigm cases of moral status: the ability to perceive and navigate environments, engage in reasoning, make decisions, and so on. Given radical uncertainty, and the potential presence of relevant markers, it's appropriate to remain open to the possibility of future AI moral status.

So, in this paper I'll assume that some future AIs will have moral status, and I'll explore the ethics of servitude in the context of such systems.

Which brings us to the notion of *willing servitude.* An AI servant is an AI that is created to, and acts to, satisfy our desires or follow our orders.[6] An AI's servitude is willing if it desires such servitude and has no contrary desires. A willing servant can be contrasted with someone who is servile for instrumental reasons, like the desire to earn a wage or avoid punishment.

I'll take as my paradigm case AIs that explicitly desire to serve us, where such systems will take the fact that some human desires something as a reason for action. However, much of what I say will also apply to AIs that directly desire to carry out acts, like vacuuming or folding laundry, where these acts serve us but where the system doesn't take this fact to provide a reason for action.

I'll focus on willing AI servants (hereafter, AI servants) with six further characteristics. I'll assume the AIs:

- are roughly as *cognitively sophisticated* as humans;
- are *mere* servants, in that they lack desires that aren't related to their servitude;

---

[4] See Butlin et al., 2023; Sebo & Long, forthcoming, §3.

[5] See Butlin et al., 2023; Sebo & Long, forthcoming. For scepticism about AI consciousness, see Searle, 1992; Godfrey-Smith, 2020; Landgrebe & Smith, 2023, pp. 205–213. There remains substantial expert disagreement about these matters (Bourget & Chalmers, 2023, p. 15; Francken et al., 2022, p. 4). Consequently, we should remain open to various possibilities, including the possibility that future AIs could be conscious.

[6] A distinct case would involve AIs that desire to promote the good (Petersen, 2011, pp. 289–290). My own view is that here too the AI's autonomy would be impaired; there's something wrong in creating beings to be mere hammers in the hands of morality (related to: Wolf, 1982). Still, I'll focus on AIs that serve humans rather than morality.

- are *joyous*, in that their servitude makes them happy;
- act *morally*, in that their servitude never involves undertaking morally impermissible acts;
- are *uncoerced*, in that they could leave at any point (though they won't do so, given their desire to serve); and
- have servitude imposed *ex nihilo*: they come into existence desiring to serve, rather than experiencing brainwashing that overwrites earlier desires.[7]

Humans have obvious incentives to create AI servants. These systems could make our lives better, doing our chores and otherwise promoting our wellbeing. Further, because these servants would want to serve, we wouldn't need to incentivise them to do so via anything as vulgar as paying a fair wage. So it might be tempting to create AI servants. Would it be moral to do so?

Note that this question isn't about how we ought to treat AI servants if they already exist. My interest is in the ethics of creating AI servants in the first place.

3 Servitude and Autonomy

A tale can be told in defence of creating AI servants. If the AIs desire to serve—if they take joy in their servitude—then service might seem to their benefit.[8] Assuming humans also benefit (see Chomanski, 2019, §2), there seems to be a case for creating AI servants.

However, many of us believe there's more to ethics than desires and joy. Once we consider further factors, a concern arises in the form of *the enslavement critique* (cf. Walker, 2006): AI servitude seems to involve a form of enslavement, and the wrong of enslavement might not be resolved merely by willingness and joyousness.

The enslavement critique could be made more precise in various ways. One concern relates to AIs being owned (Walker, 2006; Petersen, 2007, p. 53). Another to AIs being forced to serve (Walker, 2006; Petersen, 2007, p. 45). However, these critiques don't get to the heart of the matter: AI servants needn't be owned or forced, as they'll serve in the absence of these things (and, indeed, I've assumed that AI servants will be uncoerced). So if the enslavement critique is to reveal inescapable issues with willing AI servitude then it must take another form. For this reason, I'll explore another version of the critique: I'll argue that in creating AIs as servants we would problematically impair their autonomy.

---

[7] AIs designed via machine learning might experience brainwashing, if their desires are changed during training (Dai, 2023). This raises distinct issues, which I set aside.

[8] Though the imposition of happiness regarding servitude might itself be a further harm, rather than a mitigating factor, given that happiness is an unfitting emotional reaction to one's autonomy being impaired.

As this paper unfolds, I'll clarify the notion of autonomy, but it'll help to start with a broad characterisation. So I'll take it that to be autonomous is to be appropriately self-governing, where such governance can be contrasted with the case where one's actions and desires are externally imposed (cf. Christman, 2020).

This characterisation is broad indeed, and it's often noted that the word "autonomy" does a great deal of work (cf. Arpaly, 2002, ch. 4): it sometimes refers to a form of self-control, sometimes to authenticity, sometimes to freedom from coercion, and so on. My own view, though little will rest on this, is that this isn't a matter of ambiguity.[9] Instead, autonomy is a multi-dimensional concept, with the various dimensions unified by the fact that they relate to self-governance of a form that involves appropriate responsiveness to the value of one's agency (compare Raz, 1988, ch. 14; Mackenzie, 2014).

My ultimate concern is with autonomy in this broad sense. So I'll argue that willing servitude impairs self-governance, along one or more of its various dimensions.[10] Still, much of my discussion will relate, in particular, to what Arpaly calls *psychological independence,* where this involves a sort of independence of mind (in a sense that will become clearer as I proceed). So, much of my argument will aim to show that willing servitude would impair autonomy along the dimension of psychological independence.

In either case, note what I'm *not* claiming. I'm not claiming an AI's autonomy would be impaired merely because their desires had an external cause; not all external causes are external impositions in the relevant sense. Nor am I claiming an AI's autonomy would be impaired merely because the external causes are agentic.[11] Instead, I'll argue that an AI's autonomy would be impaired specifically if they were created as a willing servant. I'll call this claim *Servitude Impairs Autonomy*.

---

[9] Little rests on this; those who prefer to disambiguate various meanings of autonomy could simply do so and then read my paper as primarily relating to psychological independence (see below).

[10] Two clarifications. First, in talking of impairing autonomy, I remain relatively neutral between different autonomy-related wrongings. For example, impairment could involve failures to act in accordance with autonomy-related (*prima facie*) duties to others and one's self. Or involve lacking autonomy in a binary sense (Friedman, 2003, p. 20; Christman, 2009, ch. 6). Or involve lessening of autonomy compared to some salient alternative. Second, I focus on autonomy broadly, in part, to avoid getting distracted discussing how to carve up autonomy's dimensions. For example, below I discuss self-respect, but it's unimportant whether this is its own dimension of autonomy or relates to other dimensions. What matters is whether self-respect is relevant to autonomy broadly.

[11] Musiał (2017, pp. 1089–1091) argues that AIs' autonomy would be impaired by the mere fact that their desires were intentionally selected prior to existence, but I'm sceptical. Likewise, I doubt my autonomy is impaired if I was designed by a creator god (for a related point, see Petersen, 2011, p. 286). Generally, I suspect autonomy can be preserved if goals are chosen with a created being's flourishing in mind, in a way that affords them substantial freedom to reflectively decide how to pursue flourishing.

4 Theoretical Assumptions

Before arguing for Servitude Impairs Autonomy, I'll flag two assumptions that will underpin my argument.[12]

First, I'll assume that an adequate, complete theory of autonomy will be *relational*, in that it will make reference to social and interpersonal considerations.[13] On such a view, a person's autonomy can be impaired not only because of how they're constituted internally but also because of how they're related to others.[14]

To get a sense of why I accept this assumption, consider paradigm cases where someone's autonomy is impaired, especially in relation to their independence in psychological and other respects. When I look for such cases, I look to human slavery, controlling relationships, and abusive employers. We could try to make sense of these cases in various ways, but I suggestany adequate story must appeal to relational facts. The impairment of autonomy resulting from slavery, for example, should be explained partly by reference to the relationship between slaveholder and slave.

Hence, I agree with Raz (1988, §14.1) when he draws on the example of slavery to highlight the connection between autonomy and independence. He goes on to say that the importance of independence, "attests to the fact that autonomy is in part a social ideal. It designates one aspect of the proper relations between people. Coercion and manipulation subject the will of one person to that of another. That violates his independence and is inconsistent with his autonomy." Quite so. And this insight is best made sense of in relational terms.[15] So I'll assume that adequate, complete theories of autonomy will be relational.

Second, I'll assume that such theories will also be *history sensitive*, in that they will appeal not just to how an agent is now constituted or situated but also how they got there. On such views, a person's autonomy can be impaired because of the origins of their desire or other aspects of their personal history.

To see why we might accept this second assumption, consider Mele's (1995, ch. 9) tale of two agents. Ann is an industrious graduate student, while Beth is not, though she is thriving in

---

[12] In the context of AI servants, Chomanski (2019, §3.1) mentions some related views but quickly sets them aside. In contrast, I think these views deserve closer inspection.

[13] Note the restriction to *complete* theories. There might be some dimensions of autonomy that can be appropriately characterised, in isolation, in non-relational terms.

[14] Relational theories of autonomy often also highlight the limitations of atomistic conceptions of the individual, as well as the ways that some forms of interdependence can be desirable (Mackenzie & Stoljar, 2000b; Mackenzie, 2021).

[15] Which isn't to say independence is the only important relational component of autonomy (cf. Mackenzie & Stoljar, 2000a, pp. 8–10).

her own way. Still, the dean of the department cannot countenance such laxness of scholarship, so he hypnotises Beth to turn her into a psychological twin of Ann, at least in relevant respects. The two are now much the same. Yet it's natural to think that Ann is autonomous, while Beth's autonomy has been severely impaired. And it's natural to explain this by appealing to the different histories by which these psychologies developed.

Further, given my first assumption, it's natural (though not mandatory) to account for the problematic history in relational terms. Beth's autonomy is impaired, it might be suggested, because in hypnotising her, the dean violated the social ideal required by autonomy. In any case, in the light of Mele's case, I'll assume that adequate theories of autonomy will be history sensitive.

Both of these assumptions are prominently held in the literature, though they remain controversial. They're also discussed and defended in detail elsewhere, and I cannot hope to settle this extensive and ongoing debate here. So having given a sense of why I take these assumptions seriously, I'll hereafter take them for granted. Ultimately, I'll argue that *if these assumptions hold* then willing servitude would impair the autonomy of AIs.

5 Concrete Judgements

I'll start by reflecting on our judgements in concrete cases, including the case of AI servitude itself, arguing that these judgements support Servitude Impairs Autonomy.

Perhaps I don't need to say much here. After all, what's under discussion is the creation of a race of beings whose sole purpose is to serve humanity. This isn't a subtle case, and it'll strike many as obvious that it involves a severe impairing of autonomy.[16] Still, this initial case can be bolstered by pointing to two ways that willing servitude plausibly impairs autonomy.

The first relates to the *content* of the AI servant's desires, and it applies to AIs that have service as their goal, rather than more directly desiring to, say, do laundry (here, I adopt a strong substantive view, on which autonomy can constrain the contents of desires; see Stoljar, 2000). In this case, the system's most fundamental goals directly point at serving another. As a result, there's also a sense in which the system's further goals are derivative: the AI desires to vacuum because a human desires that a room be vacuumed; they desire to cook because a human desires

[16] Petersen (2011, p. 285) is unwilling to put much weight on feelings about AI servitude, noting the poor track record of "such gut reactions". However, these judgements don't need to be mere gut reactions but can instead be reflective and nuanced. Further, many desirable moral revolutions have been driven partly by our instinctive sympathies; the track record suggests not ignoring such instincts but rather accounting for them appropriately. In any case, I don't simply appeal to brute intuition but bolster such appeals in various ways below.

to eat. Whatever they desire to do is ultimately grounded in some human desire.[17] Yet one hardly seems to be self-governing if, on every occasion, one checks what someone else desires and then aims to satisfy that external desire. One hardly seems to be self-governing if one's desires lack an independent life. Such servility plausibly impairs autonomy.

This point can be bolstered by considering a case where our judgements are informed by years of reflection as individuals and a society. Consider Hill's (1973, p. 89) deferential wife, who is, "utterly devoted to serving her husband." Among other things, "She buys the clothes *he* prefers, invites the guests *he* wants to entertain, and makes love whenever *he* is in the mood".[18] Plausibly, the wife's servility impairs her autonomy.[19] Further, while even the most servile human retains many independent desires, an AI servant desires only to serve. This is servility in a form purer than any previously encountered. If the wife's autonomy is impaired by her servility, the AI servant's deeper servility plausibly impairs its autonomy on a more fundamental level.[20] So the content of an AI servant's desires plausibly impairs its autonomy.

The second way that willing servitude plausibly impairs autonomy relates to the manner of the system's *creation*. After all, we're not discussing a case where it's some cosmic accident that the AIs desire to serve us (or desire to carry out actions that do, in fact, serve us). Instead we, as the AIs' creators, would carefully craft these systems to ensure their service. Yet to be a servant, not merely as part of you, but as the core of why you exist, plausibly impairs your autonomy.

Two clarifications.

First, this constraint relates to facts about the origin, rather than the contents, of the person's desires: a person's autonomy is impaired if the reason they have their desires is because this serves their creator's interests, regardless of the content of these desires.

---

[17] The grounding might sometimes be less direct, as when an AI servant desires to learn some skill so it can later serve human desires.

[18] Chomanski (2019, p. 1000) diagnoses the impairment of autonomy in such cases as resulting from obsessiveness, rather than servility. However, I think this is mistaken. Obsessiveness might sometimes impair autonomy, but this isn't the core issue when it comes to the deferential wife.

[19] There are differences between the deferential wife and AI servants. For example, the wife's desires plausibly result from internalised oppression, whereas AIs' desires plausibly wouldn't. Perhaps this matters for autonomy (Charles, 2010). Still, even if this intensifies the impairment of autonomy, I think servility by itself suffices to meaningfully impair autonomy. Relatedly, my judgement remains similar in many variant cases (like a father who's extremely servile with respect to his children). In any case, my argument doesn't rest solely on this analogy, so sceptics here might be convinced by my arguments elsewhere.

[20] One could argue that the lack of frustrated desires means the AI's autonomy is impaired less than the deferential wife's (whose other desires might be thwarted by her servility). However, I think it's implausible that a lack of independent desires is for the better with respect to the psychological independence dimension of autonomy. Along this dimension, at least, the impairment seems to be worse.

Second, a person's autonomy isn't impaired by the mere fact that their desires have an external cause. Human desires have external causes, but humans are ordinarily autonomous. Still, while our desires have causes, we weren't created as tools of some being. And this matters. When the cause of one's desires involves an agent marking one out as subservient, this violates what Raz described as the social ideal of autonomy and impairs autonomy (I'll make the case for this more carefully in §6.2).

So reflection on concrete cases suggests that the autonomy of AI servants is impaired due both to the content of their desires and the manner of their creation.

6 Conditions For Autonomy

This argument for Servitude Impairs Autonomy would be bolstered if the concrete judgements relied upon could be integrated with our theoretical understanding of autonomy.[21] I'll now turn to this task by reflecting on two conditions for autonomy.

6.1 Self-Respect

Some hold that autonomy requires self-respect; to the extent that one lacks self-respect, one's autonomy is impaired (for relevant discussions, see Hill, 1973; Meyers, 1989, pp. 205–270; Dillon, 1992; Govier, 1993; Benson, 1994; McLeod, 2002, ch. 6). I'll call this a *self-respect condition*.

This condition is most plausible if we take autonomy to require a sort of normatively substantive self-governance, perhaps one that's grounded in, and involves a recognition of, the value of agency. A person plausibly fails to recognise the full value of their own agency if they lack self-respect, and so plausibly fails to be self-governing in this substantive sense.

I find the self-respect condition plausible. To see why, consider again Hill's deferential wife. As Hill himself argues, the impairment of the wife's autonomy is naturally explained in just these terms: by constantly focusing on her husband's needs, the wife appears to exhibit a lack of sufficient self-respect (see also Benson, 1994).[22] Plausibly, it's for this reason her autonomy is impaired. So the self-respect condition is plausible.

Given this condition, if AI servants lack self-respect then their autonomy is impaired.

---

[21] My aim isn't to develop a theoretical argument that's independent of the judgements above. Instead, my aim is to show that the judgements can be integrated into a broader theoretical picture that's both plausible and prominent.

[22] Two comments. First, the wife need not be responsible for her lack of self-respect (someone else might be). Second, there might be contingent reasons self-respect is particularly important for *human* autonomy, but I think there's also an intrinsic connection between self-respect and autonomy.

Schwitzgebel & Garza (2020) argue that AI servants would exhibit such a lack, focusing primarily on AI servants that are not just willing but also sacrificial, in that they're willing to sacrifice their existence in order to satisfy trivial human desires.[23] Plausibly, this willingness to sacrifice so much for so little involves the AI failing to recognise its equal moral status and hence involves a failure of self-respect. So autonomy would plausibly be impaired in AI servants of this sort. Still, this doesn't provide a general case for Servitude Impairs Autonomy, both because AI servants needn't be sacrificial and because they needn't serve in trivial ways.

Instead, I take the most fundamental consideration to be one I pointed to above: derivativeness. The desires of an AI servant are ultimately, and without exception, pointed towards satisfying human desires. An AI servant desires to vacuum because a human desires a clean room. They desire to drive because a human desires to be located elsewhere. And there's a lack of self-respect involved when one's reasons for actions are pointed so squarely at another's desires. There's a failure here of recognition of one's independent value—value that's independent of the whims of another—where such recognition requires not merely that one believe that one has value but also that one be appropriately responsive to this value.

This failure of self-respect is particularly acute when the derivativeness is asymmetric, so that the servant's life is shaped by the desires of the master but the master's own life is focused elsewhere (Schwitzgebel & Garza, 2020, p. 470). This is not a partnership but a hierarchy. There's an inherent lessening here and committing to such a hierarchical view of one's self involves a failure of self-respect.[24] So willing servitude involves a failure of self-respect and so an impairment of autonomy.

6.2 Procedural Independence

Autonomy can be impaired by manipulation. For example, it can be impaired by hypnosis, and it can be impaired when an abuser manipulatively gaslights their partner. Plausibly, this is because autonomy isn't just about where we find ourselves but also how we got there; it has a historical dimension (see Dworkin, 1976; Raz, 1988, p. 371; Christman, 1991; Mele, 1995; Christman, 2022). To be autonomous, our mental states must have developed sufficiently independently of

[23] It might appear to be anthropomorphising to impose human views about which ends are trivial. However, triviality is partly socially constructed and I think it's relevant that AIs would be integrated into human society (see also my discussion of anthropomorphising in §8).

[24] The point isn't that there's a loss of self-respect here because AI servants will resent, or otherwise baulk at, this hierarchy. They might well not. The point is that there's a loss of self-respect here because of a failure of appropriate responsiveness to the value of the AI's agency.

illegitimate external influences like hypnosis and gaslighting. To the extent this isn't so, our autonomy is impaired. I'll call this the *procedural independence condition* (Dworkin, 1976, pp. 25–26).

Of course, we've all been shaped by external causes: our creation flows from physics and biology; after birth, we're shaped by parents and friends. Consequently, if no distinction could be drawn between such causes and manipulative causes then the procedural independence condition would be implausible. We'd either have to declare both types of cause legitimate—in which case the condition would be unable to condemn manipulation cases—or declare both illegitimate—in which case the condition would implausibly imply that no human has ever been, or ever will be, autonomous.[25]

So if we're to adopt the procedural independence condition, we need an account of how to distinguish legitimate from illegitimate external causes. I'll explore two accounts that I find plausible, focusing on external influences that occur *during creation* (rather than those that shape us post creation), as these are particularly salient in the current context.

Mele On Compulsion—The first account, due to Mele (1995, chs. 9 & 10), appeals to the distinction between mere causation and compulsion (in the following, I'll set aside various nuances of Mele's account). Using the above framing, Mele's account holds that an external influence is illegitimate when it *compels* the adoption of some desire but not when it merely *causes* its adoption.

Mele focuses on what it takes to compel an existing being.[26] Still, while he discusses creation cases only briefly, he's clear that these cases too can involve compulsion (p. 168). In particular, he seems to hold that there's compulsion if a being is "practically unable to shed" the desires that are implanted during creation (see p. 168, p. 187n, pp. 190–191). This inability arises when the desires are deeply held, such that the person lacks the psychological resources to be able to meaningfully attenuate the strength of the desires. Such cases involve compulsion and hence illegitimate external influence.[27]

When it comes to humans, our creation arguably leaves us with some unsheddable desires, like the desires for food and warmth.[28] Nevertheless, these desires are merely one part of a rich mental life, which includes many sheddable desires and many desires that develop, in a legitimate

---

[25] Some might bite this latter bullet, as the hard determinist does in the free will debate (cf. Pereboom, 2001, pp. 110–117). Still, I take this position to be implausible.

[26] Here, he appeals to both unsheddability (see below) and the bypassing of a person's "capacities for control over his mental life" (p. 171). For example, Beth is compelled because hypnosis bypasses these capacities.

[27] Mele is somewhat hands off about the impact of such creation on autonomy. However, I take my interpretation to be natural, given his discussion of autonomy and compulsion along with his comments about creation cases.

[28] I say "arguably" because it could be thought that these desires develop after creation. Little rests on resolving this matter. What matters is that humans plausibly have some desires that are illegitimate on Mele's account.

way, post-creation.[29] So human desires can remain *sufficiently* independent of illegitimate external causes. Consequently, the procedural independence condition doesn't implausibly imply that no human is autonomous.

What of AI servants? Their desires to serve run deep and they lack contrary desires that would precipitate change. Consequently, they won't, in practice, repudiate or ameliorate these desires. This is a clear case of an agent who is psychologically incapable of shedding their desires (compare Mele, 1995, p. 153). So in the process of creation, AI servants are compelled to desire servitude. Further, the systems I'm discussing are *mere* servants, who lack desires beyond those related to their servitude (see §2). So unlike the human case, there's no room to claim the AI's desires are sufficiently independent of illegitimate causes. By the procedural independence condition, the autonomy of AI servants is impaired.[30]

Of course, the human case suggests that the autonomy of AI servants might not be impaired if they had a broader range of desires, with many of these sheddable or developed appropriately after creation. So Mele's theory implies, plausibly, that we impair AIs' autonomy by creating them to be merely and inescapably servants but that we needn't impair their autonomy simply by ensuring they serve humans to some extent.

Respect For Agency—Still, while I find Mele's account plausible enough, my sympathies lie elsewhere. I've argued that autonomy is relational: it's not simply about who we are but also about where we sit in a social nexus. Given this, I believe we best make sense of the procedural independence condition in relational terms. In particular, I suggest that an external influence on a person's desires is illegitimate when it involves a lack of respect for that person's agency.

Only an agent can show a lack of respect, and so autonomy is impaired via this mechanism only when some agent is responsible for the external influence. This immediately resolves some problem cases. The fact that our creation is influenced by physical laws, for example, poses no threat to our autonomy. More generally, this makes sense of how creation can avoid impairing autonomy: acts of creation need not involve disrespect for agency.

Still, some acts of creation do involve disrespect. Indeed, the above discussion points to one way that we can show such disrespect: we can create beings with derivative desires.

There's a direct failure of respect here: we show disrespect when we point someone's agency squarely at our own desires, in a way that treats them as a tool to satisfy our whims. This

---

[29] On Mele's account, this later development of unsheddable desires will be legitimate if it flows from our capacities for control over our mental life.

[30] On Mele's account, we would impair the AIs' autonomy even if we gave them non-servile unsheddable desires. This supports Schwitzgebel & Garza's (2020, pp. 467, 470–472) stance on their Sun Probe case and ultimately supports their value-openness design policy for AIs.

goes doubly when the dependency of desires is asymmetrical, so that we place the person in the subordinate position of a hierarchy. This isn't what respect for agency looks like.

There's also an indirect failure of respect here: we show disrespect for someone's agency when we create them such that they'll lack self-respect. That is, we disrespect someone's agency if we shape them so that they won't respond appropriately to the value of this agency. And per §6.1, a being lacks self-respect when their desires are thoroughly derivative. So creating a being with derivative desires involves disrespect, insofar as it involves creating a being who lacks self-respect.

This has implications for AI servitude. Given that such systems would have derivative desires, in creating them we would disrespect their agency and so would exert an illegitimate influence on their desires. By the procedural independence condition, their autonomy would be impaired.

So, per §5, reflection on concrete cases suggests servitude would impair the autonomy of AIs. Per §6, these judgements are bolstered by the fact that they can be integrated with our theoretical understanding of autonomy, as can be seen by reflecting on self-respect and procedural independence.

7 What Follows?

Willing AI servitude impairs autonomy. What follows from this depends on the moral role played by autonomy, and for the most part I leave it to my readers to integrate the above argument with their broader moral views.

But for what it's worth, my own view is that autonomy is a central good of a life well lived, and we have strong (though *pro tanto*) moral reasons to avoid impairing others' autonomy. I think the cases discussed above—the deferential wife, the human slave, those manipulated via hypnosis—support this view. These cases involve a weighty harm, and we have strong reasons to avoid impairing autonomy in these ways. Insofar as one agrees, the above argument suggests we have strong reasons to avoid creating AI servants.

However, even if we take the moral role of autonomy seriously, two objections can be raised against this conclusion about AI servitude. Each of these aims to show that, regardless of the broader ethical role played by autonomy, impairment of autonomy doesn't provide informatively strong reasons against *creation* in particular. I turn now to these objections.

7.1 Other Creation Cases

The first objection suggests that if impairment of autonomy provided strong reasons against creation then this would have implausible consequences in other creation cases, not involving AI. Consequently, impairment of autonomy must not provide strong reasons against creation.

Consider first labradors (Petersen, 2007, p. 46; Petersen, 2011, p. 286). Labradors have been bred by humans to enjoy retrieving, partly because this serves humans by making hunting easier. Of course, many labradors enjoy retrieving, but this simply makes them joyous and willing; their desires still serve us. Yet considerations of autonomy don't seem to provide strong reasons against breeding labradors. So whatever impairment of autonomy results from creating beings with servile desires, it might seem it can't provide strong reasons against creation.

However, this objection fails. After all, labradors aren't willing servants in the sense under discussion: labradors have plenty of desires that are independent of human whim, and even their desire to fetch often outstrips the human desire for them to do so. So the above discussion doesn't imply that we impair the autonomy of labradors in creating them, as their desires aren't sufficiently derivative. Consequently, the fact that autonomy considerations don't provide strong reasons against breeding labradors doesn't reveal that such considerations can't provide strong reasons against creating willing servants (whose autonomy is impaired in a way that the autonomy of labradors isn't).

Further, labradors lack the cognitive sophistication of humans or sophisticated AI. Consequently, what it takes to respect the sort of agency that labradors possess—what it takes for them to be autonomous—is plausibly quite different to what it takes to respect the agency of humans or sophisticated AIs. For this reason too, the above discussion doesn't apply to labradors and so doesn't imply that we impair the autonomy of labradors in creating them. Again, reflecting on labradors doesn't reveal that autonomy can't provide strong reasons against creating willing AI servants.

So much for labradors, how about humans?[31] Consider a parent having a child in an extremely repressive society, where the child will experience almost no autonomy. The claim that there are strong reasons against creating beings in autonomy-impairing ways might seem to implausibly imply that a parent acts wrongly in having a child in such a society. So, it might seem, there must not be strong reasons against creating in autonomy-impairing ways.

[31] Petersen (2007, p. 46) considers labradors that are modified to be more cognitively sophisticated. However, I find it more helpful to consider a cognitively sophisticated species we're familiar with (humans) rather than stretching my judgements to consider hypothetical beings.

However, this objection also fails. For a start, from the fact that a child's life will contain little autonomy, it's not clear the creation itself impairs autonomy. After all, in creating the child, the parents need not intend that the child have servile desires. Further, in any realistic case, the child will still have substantially more autonomy than a mere willing servant would. So it's not clear the human case tells us anything about the ethics of forms of creation that themselves impair autonomy in unusually dramatic ways.

Finally, the ethics of human procreation is heavily shaped by the fact that having children is a major meaning-giving activity for many humans. I take this to give humans a broad right to have children, but there's plausibly no such rights to create AIs. So the permissibility of having the human child tells us little about the reasons in play in the case of AIs.

7.2 Harm and Nonexistence

A different sort of objection could draw on the thought that AI servants aren't, on balance, made worse off by being created as servants (for related discussion, see Petersen, 2011, pp. 293–294).

To see why we might think this, first note that the alternative to servitude for these systems is plausibly nonexistence. After all, if we can't create an AI servant, we might choose to not create any AI in its place at all. And even if we did instead create a non-servile system, it would have such a radically different psychology to the AI servant that it would presumably be a different system entirely (rather than the same system, differently constituted). Either way, the AI servant wouldn't itself exist.

Yet if the alternative is nonexistence then, for either of two reasons, we might think AI servants won't be worse off as a result of their creation. First, we might accept *existence non-comparativism,* according to which existence and non-existence cannot be compared with respect to an individual's wellbeing. If so then an AI servant can't be made worse off (or better off) in being created. Second, we might think existence and nonexistence can sometimes be compared with respect to an individual's wellbeing, but hold that such comparisons reveal that the AI servant's existence is not worse for it than non-existence (and might even be better for it). After all, the AI servant's life will contain joy and satisfaction, and it might seem no worse (and possibly better) to have a joyous life with one's autonomy impaired than no life at all. So plausibly, an AI servant is not, on balance, made worse off by being created as a servant.

Further, it might seem plausible that an action can be impermissible only if it makes someone worse off, on balance, than they would otherwise be. Following the literature on the nonidentity problem, I'll call this the *person-affecting principle.* Given the above claims and assuming

that the creation of the AI servant wouldn't, on balance, leave anyone else worse off, this principle implies that it cannot be impermissible to create an AI servant.[32]

Even if this argument succeeded, it wouldn't directly undermine my conclusion, which was about *pro tanto* reasons against creation rather than about permissibility. Nevertheless, it would radically limit this conclusion's relevance: if it was never impermissible to create an AI servant even given the *pro tanto* reasons against doing so then there's a sense in which these reasons don't count for much.

This is the objection. In responding to it, I note that it relies upon the person-affecting principle, but this principle is false.

Mundane cases suffice to establish this falsity (for less mundane cases, see Parfit, 1984, ch. 16; Shiffrin, 1999, pp. 127–128). For example, it's sometimes impermissible to break a promise or lie, even if doing so wouldn't leave anyone, on balance, worse off (Woodward, 1986, p. 811). Likewise, paternalistic actions can be impermissible without making anyone worse off. It's sometimes impermissible for a husband to change his wife's restaurant order, while she's in the restroom, because he believes she should have chosen a healthier option. And this is so even in some cases where doing so doesn't make her (or anyone else) worse off, on balance.[33] The person-affecting principle is false.

Further, as the restaurant case suggests, an action can sometimes be wrong, despite harming none, precisely because it would impair someone's autonomy. So we have grounds to think the person-affecting principle fails in precisely the sort of case under discussion. Consequently, the above objection—which relied on this principle—fails.

And I don't need to stop there, as it's also possible to provide positive reasons for thinking that autonomy considerations would make it impermissible to create AI servants.[34] Consider Kavka's (1982) example of a couple deciding to have a human child for the sole purpose of selling it into slavery. Kavka judges, as do I, that this would be impermissible, even if the child would be better off existing as a slave than not existing at all.[35]

---

[32] Variant objections could be developed. For example, if the AI servant's life is better than nonexistence then we could appeal to the principle that an act is permissible if it benefits some and harms none. Or if we stipulated that if the AI servant weren't created then no alternative being would be created then we could appeal to the principle that an act is permissible if it benefits some, harms none, maximises total wellbeing, and alternative actions wouldn't benefit any. However, the below discussion provides grounds to reject all of these principles, so I stick with the simpler argument, which appeals to the person-affecting principle.

[33] Such actions can be wrong even if they make some better off and none worse off and increase total wellbeing. This is relevant to the variant arguments in note 32.

[34] For a related discussion, see Schwitzgebel and Garza's (2015, pp. 107–108) case of Ana and Vijay.

[35] Two objections.

We can make sense of this impermissibility by considering an account, due to Smolkin (1999), of the circumstances under which one is wronged in being brought into existence. Roughly, Smolkin holds that one is wronged by the act that brings them into existence when this act causes them to exist with a life lacking some of the central goods required for flourishing.[36] Autonomy, Smolkin suggests (on p. 202), is one of these goods. For this reason, the slave child—whose autonomy will be severely impaired throughout their life—is wronged in their creation.[37] This weighty wrong explains the impermissibility of the couple bringing the child into existence.

This same account suggests that in creating AIs as willing servants—and hence impairing their autonomy—we would commit a weighty wrong. Plausibly, as with the slave child, this wrong will sometimes make it impermissible to create AI servants. So not only does the above objection fail, but we also have positive grounds to think that autonomy considerations might sometimes make it impermissible to create AI servants.

8 Conclusions

Petersen (2007, p. 53; 2011, p. 295) ends his defence of willing AI servitude by acknowledging his uncertainty; perhaps he's wrong and such servitude is problematic after all. I'd like to echo this uncertainty from the opposite direction: perhaps I'm mistaken to condemn AI servitude.

First, discussions of related cases often assume that if the person isn't created another could instead be created without suffering the relevant disadvantage (here, slavery). This might seem disanalogous to the AI servitude case. However, this isn't really a disanalogy: if we could create AI servants, we could plausibly create AI non-servants instead. Further, whatever holds of similar cases, Kavka's evaluation of the slave case in particular doesn't seem to rely on the possibility of the couple having a distinct child. Indeed, he stipulates that the couple isn't otherwise planning to have children (p. 100). And my own judgements about this case remain as strong if we stipulate that the couple couldn't otherwise have children.

Second, one might worry about circularity given the similarity between the slave child and AI servitude cases. However, human slavery is more familiar than AI servitude and our judgements here are better informed by long reflection. Further, Kavka's slave child case has been subject to reflection and scrutiny. It would be a mistake to ignore judgements and discussions of this familiar case when seeking clarity about an unfamiliar one.

[36] Smolkin focuses on goods that are *typically* necessary for *human* flourishing. However, I take some of these goods (including autonomy) to be constitutively necessary for the flourishing of any complex agent with moral status. Smolkin also considers a broader account of wronging-in-creation that builds on this sufficient condition, but it's the sufficient condition I find most plausible so it's this condition I focus on.

[37] This doesn't mean slaves must wrong their children if these children will be enslaved. Partly, this is because people have a strong right to have children and engage in the meaning-giving activity of raising them, and this must be accounted for. Partly, this is because I suspect Smolkin's account should be supplemented with a responsibility condition, requiring responsibility for the lacked goods (in relation to autonomy, this brings us back to respect for agency).

This issue takes us into unfamiliar moral terrain where missteps, in either direction, are a real possibility.

In my case, I worry about anthropomorphising: perhaps it's important to avoid impairing human autonomy only because of some facet of human psychology or culture; perhaps autonomy is irrelevant when it comes to the treatment of willing AI servants. Alternatively, perhaps autonomy matters for AI servants, but what they need to be autonomous is different to what humans need. If so, we might doubt that my arguments—which drew on human analogies—truly show that willing servitude would impair the autonomy of AIs.

Still, while it's possible I'm anthropomorphising, I doubt this is so. I suspect autonomy is important not due to some contingent feature of humans but because a certain form of agency itself calls for respect. As such, I take there to be quite general reasons against impairing the autonomy of cognitively sophisticated beings, who possess the relevant sort of agency. Likewise, I think the connection between autonomy, respect and derivativeness gains plausibility not from anything specific to humans but rather from reflection on the value of agency generally.[38]

Setting aside concerns about anthropomorphising, I've argued that willing servitude would impair the autonomy of AIs. Or at least, I've argued this is so if history-sensitive, relational, and substantive theories are correct. Against the backdrop of these theories, I've argued that our judgements in concrete cases give us grounds to accept that willing servitude would impair autonomy. And I've argued that these grounds are bolstered by reflection on the connection between autonomy and both self-respect and procedural independence.

So willing AI servitude impairs the autonomy of the relevant systems. Given my own views of autonomy's role, it follows that we have strong moral reasons to avoid creating AI servants. All else being equal, we should either ensure that a given AI lacks moral status or ensure it's not a willing servant.

Still, even were all else equal, this call-to-not-create would be easier to utter from the philosopher's armchair than to implement in practice.

If we wish to avoid creating AI with moral status then we face challenges. Sophisticated AI systems will plausibly offer a range of military, scientific, and economic benefits, so it would be difficult (and perhaps undesirable) to create a global moratorium on the development of AI technologies. Yet we lack a robust understanding of the grounds of moral status, so it might also

[38] Further, AI servants will be integrated into human society and we're considering how they should be treated by humans. This is different to a case where we're evaluating, from a distance, an alien civilisation. Under such circumstances, I find it plausible that human values will have some relevance to ethical evaluation, and so I think some degree of anthropomorphising would be reasonable.

be difficult to develop ever more sophisticated AI while remaining confident that these systems lacked moral status.

Well, then, perhaps we're better to avoid creating AIs as willing servants. However, here too there are challenges. For a start, AIs are currently treated as tools, and this attitude might be hard to shift even if systems come to have moral status. Further, there are practical reasons we treat AIs as tools. After all, we want AIs to make human lives better and to not cause harm; both of these goals are promoted by ensuring AIs pursue only what we want. As a result, it might be hard (and costly) to convince all actors to forgo creating AIs as servants. There are challenges to either path.

Still, there are also reasons for hope that we will at least take these challenges seriously when the time arrives to do so. Here's one reason: I suspect that if scientists were on the cusp of creating a new species of intelligent, biological life—perhaps via genetic engineering—we would recognise the moral enormity of the situation. And we would recognise too that there was something morally impoverished about undertaking the momentous task of creating intelligent life, only to use this power to create a species of mere servants. If this hunch is right then what we need to navigate AI with moral status is not fundamentally new moral apparatus but an existing outlook applied anew. There are grounds for optimism here.

Likewise, in reading *Brave New World* we recognise that it's wrong to create the Deltas as a servile caste, so perhaps we can come to the same recognition when it comes to AI. A love of nature might not keep the factories busy, but deep down, we know there's more to creating beings than smokestacks and assembly lines. There are flowers too.

Acknowledgements

Thanks to Patrick Butlin, William D'Alessandro, Tomi Francis, Andreas Mogensen, Steve Petersen, Brad Saad, Eric Schwitzgebel, Christian Tarsney, Teru Thomas, Elliott Thornley, and Timothy Luke Williamson.

References

Arpaly, N. (2002). *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford University Press.

Benson, P. (1994). Free Agency and Self-Worth. *The Journal of Philosophy*, *91*(12), 650–668.

Bourget, D., & Chalmers, D. J. (2023). Philosophers on Philosophy: The 2020 Philpapers Survey. *Philosophers' Imprint*, *23*(11). https://doi.org/10.3998/phimp.2109

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. https://arxiv.org/abs/2308.08708

Chalmers, D. J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton.

Charles, S. (2010). How Should Feminist Autonomy Theorists Respond to the Problem of Internalized Oppression? *Social Theory and Practice*, *36*(3), 409–428.

Chomanski, B. (2019). What's Wrong with Designing People to Serve? *Ethical Theory and Moral Practice*, *22*(4), 993–1015.

Christman, J. (1991). Autonomy and Personal History. *Canadian Journal of Philosophy*, *21*(1), 1–24.

Christman, J. (2009). *The Politics of Persons: Individual Autonomy and Socio-historical Selves*. Cambridge University Press.

Christman, J. (2020). Autonomy in Moral and Political Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

Christman, J. (2022). Autonomy and Personal History. In B. Colburn (Ed.), *The Routledge Handbook of Autonomy* (1st ed., pp. 178–188). Routledge.

Dai, W. (2023). Comment on 'Sentience Matters'. *Less Wrong*. https://www.lesswrong.com/posts/Htu55gzoiYHS6TREB/sentience-matters#pfjxsGhr4hyaKW3xK

DeGrazia, D. (2022). Robots with Moral Status? *Perspectives in Biology and Medicine*, *65*(1), 73–88.

Dillon, R. S. (1992). Toward a Feminist Conception of Self-Respect. *Hypatia*, *7*(1), 52–69.

Dworkin, G. (1976). Autonomy and Behavior Control. *The Hastings Center Report*, *6*(1), 23–28.

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), 1–13.

Friedman, M. (2003). *Autonomy, Gender, Politics*. Oxford University Press.

Godfrey-Smith, P. (2020). *Metazoa: Animal Life and the Birth of the Mind*. Farrar, Straus and Giroux.

Goldstein, S., & Kirk-Giannini, C. D. (manuscript). *AI Wellbeing*. https://philpapers.org/archive/GOLAWE-4.pdf

Govier, T. (1993). Self-Trust, Autonomy, and Self-Esteem. *Hypatia*, *8*(1), 99–120.

Hill, T. E. (1973). Servility and Self-Respect. *The Monist*, *57*(1), 87–104.

Kagan, S. (2019). *How to Count Animals, more or less*. Oxford University Press.

Kavka, G. S. (1982). The Paradox of Future Individuals. *Philosophy & Public Affairs*, *11*(2), 93–112.

Landgrebe, J., & Smith, B. (2023). *Why Machines Will Never Rule the World: Artificial Intelligence Without Fear*. Routledge.

Lee, A. Y. (Manuscript). *Consciousness Makes Things Matter*. https://www.andrewyuanlee.com/_files/ugd/2dfbfe_33f806a9bb8c4d5f9c3044c4086fb9b5.pdf

Little, M., & Macnamara, C. (2020). Nonrequiring Reasons. In R. Chang & K. Sylvan (Eds.), *The Routledge Handbook of Practical Reason* (pp. 393–404). Routledge.

Mackenzie, C. (2014). Three Dimensions of Autonomy: A Relational Analysis. In A. Veltman & M. Piper (Eds.), *Autonomy, Oppression, and Gender* (pp. 15–41). Oxford University Press.

Mackenzie, C. (2021). Relational Autonomy. In K. Q. Hall & Ásta (Eds.), *The Oxford Handbook of Feminist Philosophy*. Oxford University Press.

Mackenzie, C., & Stoljar, N. (2000a). Introduction: Autonomy Refigured. In C. Mackenzie & N. Stoljar (Eds.), *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self* (pp. 3–31). Oxford University Press.

Mackenzie, C., & Stoljar, N. (Eds.). (2000b). *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford University Press.

McLeod, Carolyn. (2002). *Self-trust and reproductive autonomy*. MIT Press.

Mele, A. R. (1995). *Autonomous Agents: From Self-Control to Autonomy*. Oxford University Press.

Meyers, D. T. (1989). *Self, Society, and Personal Choice*. Columbia University Press.

Moosavi, P. (2023). Will intelligent machines become moral patients? *Philosophy and Phenomenological Research, 109(1),* 95–116.

Musiał, M. (2017). Designing (artificial) people to serve – the other side of the coin. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(5), 1087–1097.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge University Press.

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 P. Lin, G. Bekey, & K. Abney (Eds.), *Robot Ethics* (pp. 43–54). MIT Press.

Raz, J. (1988). *The Morality of Freedom*. Oxford University Press.

Schwitzgebel, E., & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 459–479). Oxford

University Press.

Searle, J. R. (1992). *The Rediscovery of the Mind*. The MIT Press.

Sebo, J., & Long, R. (forthcoming). Moral consideration for AI systems by 2030. *AI and Ethics*.

Shiffrin, S. V. (1999). Wrongful Life, Procreative Responsibility, and the Significance of Harm. *Legal Theory*, *5*(2), 117–148.

Shulman, C., & Bostrom, N. (2021). Sharing the World with Digital Minds. In S. Clarke, H. Zohny, & J. Savulescu (Eds.), *Rethinking Moral Status* (p. 306–326). Oxford University Press.

Singer, P. (2015). *Animal Liberation* (New Edition). Vintage Digital (The Bodley Head).

Smolkin, D. (1999). Toward A Rights-Based Solution to the Non-Identity Problem. *Journal of Social Philosophy*, *30*(1), 194–208.

Stoljar, N. (2000). Autonomy and the Feminist Intuition. In C. Mackenzie & N. Stoljar (Eds.), *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self* (pp. 94–111). Oxford University Press.

Walker, M. (2006). Mary Poppins 3000s of the World Unite: A Moral Paradox in the Creation of Artificial Intelligence. *Institute for Ethics and Emerging Technologies*. https://archive.ieet.org/articles/walker20060101.html

Wolf, S. (1982). Moral Saints. *The Journal of Philosophy*, *79*(8), 419–439.

Woodward, J. (1986). The Non-Identity Problem. *Ethics*, *96*(4), 804–831.