Journal of Libertarian Studies
Volume 27, No. 1 (2023): 62–86

Should the State Prohibit the Production of Artificial Persons?

Bartlomiej Chomanski

ABSTRACT: This article argues that criminal law should not, in general, prevent the creation of artificially intelligent servants who achieve humanlike moral status, even though it may well be immoral to construct such beings. In defending this claim, a series of thought experiments intended to evoke clear intuitions is proposed, and presuppositions about any particular theory of criminalization or any particular moral theory are kept to a minimum.

Treating modern-day artificially intelligent agents as little better than obedient servants raises few eyebrows. This seems unobjectionable enough—surely, the Siris and Alexas are not harmed by being interacted with in this manner. However, as research in artificial intelligence (AI) progresses, future artificially intelligent agents could begin to acquire properties that could endow them with a moral status of some sort. Indeed, it may become feasible to build artificial entities with the moral status of a person (Pollock 1989).

This article argues that criminal law should not prevent the creation of artificially intelligent servants who achieve humanlike moral status—except, perhaps, in cases of engineering extremely servile beings—even though it may well be immoral to construct

Bartlomiej Chomanski (bc02@amu.edu.pl) is an assistant professor of philosophy at Adam Mickiewicz University in Poznan, Poland.

The author wishes to thank Cansu Canca, Sven Nyholm, and the audiences at Northeastern University, Harvard, and TU Eindhoven for useful feedback on a previous version of the paper. The reviewer for this journal has also provided a number of very helpful suggestions and comments, which are gratefully acknowledged.

such beings. In defending this claim, it relies on a series of thought experiments intended to evoke clear intuitions, without presupposing the truth (or falsity) of any particular theory of criminalization or any particular moral theory.

While the topic may strike the reader as fanciful, it is perhaps less speculative than one might think. Recent years, after all, have seen the well-publicized controversy over whether Google's large language model LaMDA is conscious and deserves legal protection and representation (Davis 2022). Though most have found the ascription of sentience to the model unjustified, the whole affair may be seen as a harbinger of things to come as the models become more advanced and, dare one say, humanlike.

ARTIFICIAL PERSONS: MORALITY AND CRIMINAL LAW

Consider an artificially intelligent agent equipped with human-level capacities, resulting in human-level (or above) performance on a broad variety of tasks. It is possible that such agents will be created. Suppose, then, that the AI will possess whatever is required for humanlike moral status (rationality, phenomenal consciousness, possession of a moral character, or perhaps the capacity for participation in caring and/or social relations). In other words, it will be an artificial person (AP).

Suppose a programmer builds the AP in such a way that the AP's strongest desire is to serve the needs of human beings (assume that the needs are morally unobjectionable and that the manners in which they are served are also morally unobjectionable). Suppose the AP's desire to serve is extremely strong, though defeasible, and that the AP can overcome it only with great effort—Steve Petersen (2011; 2017) in his pioneering discussion of servile APs compares the strength of their desire to serve with a human being's desire to eat and drink. Moreover, suppose that fulfilling this desire brings the APs enormous satisfaction.

Is it immoral to build such beings? And if it's morally wrong to build AP servants, should it be illegal to do so? This article will take up the second question and answer it in the negative. It will argue that even though there may well be good reasons to think that it is wrong to build AP servants, the law should permit it. This is because the wrongness of building AP servants does not provide sufficient reason to criminalize the practice.

The literature on the ethics of servile APs is scarce, but a number of principal reasons to think that building servile APs is morally wrong have been identified (either explicitly when considering the question of APs or in a way that can be applied to this question): (1) it harms the APs by objectionably interfering in their autonomy (this objection is articulated and defended by Mark Walker [2006] and more recently by Maciej Musiał [2017], and rejected by Petersen [2011][1]); (2) it harms the APs by depriving them of the opportunity to fulfill their potential (articulated and disputed by Petersen [2011]); (3) it indicates an objectionable attitude (Aristotelian vice) of manipulativeness on the part of the programmers (Chomanski 2019); (4) it implicitly denies moral status to APs (compare Duff [2014]) or promotes engagement in worthless pursuits (compare Wall [2013]); (5) it brings into being entities that, because of their extreme servility, show a radical lack of self-respect (Schwitzgebel and Garza 2020); and (6) it may violate the APs' rights (compare Feinberg [1994]).

The remainder of this article argues that none of these kinds of wrongdoing provides a conclusive reason for criminalizing the practice of building servile APs (with the exception of extreme servility, which may necessitate instituting some prohibitions, though in a highly decentralized fashion). This will remain true, moreover, largely independently of the theory one adopts with regard to the relationship between morality and criminal law. Whether one operates merely on a Millian harm principle or embraces a more expansive view of the criminal law's function (legal moralism, for example), the conclusions stand.

First, for reasons explored in depth in the literature on the nonidentity problem (see, e.g., Boonin [2014], but also Gardner [2015] and, of course, Parfit [1984]), it is not obvious whether anyone is harmed when a new AP is programmed with servile desires, provided its subsequent existence is worth living. In fact, Petersen's (2011; 2017) defense of the claim that the APs are *not* harmed by being brought into existence seems quite convincing. Assuming Millian liberalism about the law (Mill 2015), if no one is harmed by an action, the state should not criminalize it. Thus, if the APs are not harmed, there is no reason from the liberal perspective to ban creating them.

---

[1] Petersen only responds to Walker, as Musial's paper is more recent. For a short critique of Musial, see Danaher (2019).

In summary, a simple argument against the criminalization of building willing servile APs presents itself:

1. The only purpose of criminal law is to prevent harm to another person.
2. Neither servile APs nor anyone else is harmed by servile APs' being brought into existence.
3. Therefore, bringing servile APs into existence should not be criminalized.

One could reject this argument, of course. One could deny premise 1, espousing a more expansive theory of what actions warrant state punishment. This objection will be addressed in the section "Moralism about Criminal Law and the Creation of AP Servants." The seemingly most promising strategy for those who wish to criminalize the creation of AP servants is to adopt moralism about criminal law (it is difficult to see a paternalistic justification for banning the creation of AP servants). Very roughly, legal moralism claims that some actions that are not harmful in Mill's sense still warrant criminal punishment due to being immoral (see Patrick Devlin [1965]) for the locus classicus of this approach). One could also reject the argument by denying premise 2. As mentioned earlier, on some views, servile APs are harmed by being brought into existence. This objection will be addressed in the following section.

Nonetheless, this article will argue that the falsity of premise 1 and/or premise 2 is still compatible with the truth of the conclusion (3).

THE HARM PRINCIPLE AND AP SERVANTS

Suppose first that, despite Petersen's claims, AP servants *are* harmed by being brought into existence. Should their production be banned in such a case? This section argues that the creation of AP servants should not be prohibited on harm grounds.

Dictating Life Plan

There seem to be two promising ways to develop the suggestion that AP servants could be harmed by being brought into existence. The first is captured by Walker's (2006) already-mentioned claim that in bringing servile APs into existence one harms them by

dictating their life plans to them. Inflicting this sort of harm on another person—the harm of being subjected to a life plan of another's choosing—while morally wrong, does not seem to be an immediate candidate for criminalization. Some manipulation of this sort may need to be criminalized, but some may not need to be. For example, suppose a parent is raising her child to be an academic superstar or a concert pianist in such a way that the child is not in a position to easily and without incurring much cost pick another path in life. While potentially immoral and harmful, this is not the kind of act that warrants criminalization. On the other hand, raising someone to be an alcoholic might be.

An appeal to an analogy might help with the verdict here.

> UNIVERSITY 1 (U1): A university professor has a very good student who comes to him for career advice. The student has excelled in his classes, as in all the others. She's majoring in philosophy, and she wants to know what sort of career she could pursue with that degree. As a philosopher, he would really like her to become his intellectual protégé and heir. This is because he finds her to be singularly well equipped to expand his own view in a new and exciting direction, in a way that he couldn't do himself (it requires expert knowledge of some empirical matters that he is unwilling or unable to master but that she knows very well). So the professor advises that she go to graduate school in philosophy. He knows she'll do well there and have a good career as an academic. But he neglects to mention any other possible career choice that he knows would suit his student equally well. He conveniently "forgets" that he knows many philosophy majors of comparable talent who went on to pursue very fulfilling careers in business, law, or journalism. He does mention some that tried these other careers and didn't find satisfaction in them. All told, he does such a convincing job that after the meeting with him, the student develops a lasting and overwhelming desire to become a philosophy PhD and a career philosopher afterwards. She trusts the professor's judgment completely and will dismiss contrary career advice from any other source.

To be sure, there is a degree of manipulation involved here. The professor has sculpted the student's life-guiding desires in an objectionable way. Put another way, it looks as though he has dictated her life plan to her. But surely this should not be a matter of criminal law. Giving misleading career advice, even to those who trust the adviser completely and are extremely likely to follow the adviser's advice, should not generally warrant criminal punishment. Thus, dictating a life plan to another person without much regard for what the person would want should not, barring special circumstances, warrant criminal punishment either.

Another analogy may help show why this is so. Suppose that, upon hearing of the professor's meeting with the student, another person, Clancy, kidnaps the professor from his office and locks him in a utility closet for some time. Such a response seems wildly excessive and unwarranted; this is partly because of its coercive, violent, and threatening nature. By the same token, then, criminalizing the professor's behavior in U1—bringing the coercive apparatus of the state to impose punishment on him—appears likewise excessive and unwarranted. For this reason, criminalizing this kind of act should be resisted.[2]

However, one may think that U1 is disanalogous to creating servile APs, holding that the following, arguably closer, analogy demonstrates that punishing the creators of servile APs would be justifiable.

> UNIVERSITY 2 (U2): The situation is as in U1, except that rather than using words to manipulate the student to dictate her life plan to her, the professor slips a generally effective mind-control pill into her glass of water. The pill works with the same final effect as the professor's words did in U1.

It may be intuitive to endorse a violent reprisal against the professor in this amended case. However, supposing this is so, the question arises of whether U2 is a closer analogy to the servile AP case than its predecessor, U1.

There are reasons to resist this claim. For starters, the mind-control pill in U2 involves *an additional element of deception*. The professor is using a deceptive means (*hiding* the pill in the student's water) rather than conducting his manipulation in view of the student. Another difference between U2 and U1 is that in U1 the manipulative interference consists in engaging with and exploiting a weakness in the victim's existing reasoning capacity and/or existing value system. However, the professor leaves the student's *existing* reasoning capacity/value system *intact*; he doesn't change anything about it or about the student's values and preferences, nor does he override her desires and values. Instead, the professor appeals to and engages with the student's current values and desires in a manipulative way.

---

[2] Perhaps there is a relevant moral difference between Clancy's actions and what the state does, but even if Clancy were acting in his capacity as a legitimate enforcer of the laws in a liberal democratic state (a police chief, for example), his acts would nonetheless continue appearing excessively violent.

In U2, by contrast, the manipulative interference consists in overriding the student's existing decision-making capacities themselves. The student's capacities have a natural history of development along a certain trajectory that is largely independent of the professor. Overriding/ignoring these capacities appears worse than merely exploiting a weakness in them. It is a violation of the student's integrity as a rational being, an assault on the mental capacities she possesses *independently of the professor's actions*.

To see how much worse the U2 type of interference is, imagine that the professor surreptitiously slipped the student a generally effective bias-correcting pill (or some other pill that improves critical thinking, or even a high-IQ or high-motivation pill), thus strengthening her reasoning capacity, in order to then convince her to go to graduate school. This still seems more wrong than exploiting a weakness in her reasoning by manipulating her into some course of action as in U1.

One might think that the preceding considerations show that U2 is a closer analogy to the design of AP servants than U1 is—after all, the programmers are the ones building the AP's reasoning system. But this is a superficial similarity. The crucial point to consider is that the designers of APs, whatever they program into their creations, are not in a position to *interfere with extant capacities of APs* that have developed independently of the programmers' actions. APs, prior to being programmed, have no nature that can be interfered with, no capacities or values that can be usurped or overridden. As a result, the question of interference does not arise. Since there's no interference in their extant capacities and value systems, the designers' decision to program the APs to be servile cannot be analogous to the professor giving the student a mind-control pill in U2.

In sum, while merely dictating a life plan to another need not justify coercive punishment, the method whereby one dictates the life plan seems to be relevant. Directly overriding a person's existing decision-making capacities, values, and preferences appears worse than exploiting a known weakness in these capacities. But such overriding does not—cannot—occur when APs are programmed to be servile. Thus, designing servile APs is relevantly unlike U2.

Nonetheless, just because U2 is not analogous to the servile AP case doesn't mean U1 is. Could there be a case which approximates more closely the process of dictating a life plan to the blank minds of APs than U1 does? Consider the following:

> UNIVERSITY 3 (U3): The situation is as in U1, except the student meets with the professor at the start of freshman year, when she has no idea what career she wants after graduation. She is entirely open to any career possible, as she has no deeply held values or preferences relevant to picking one over another. The professor immediately recognizes that the student is intellectually well equipped to advance his own view and decides to instill in her a certain range of values (love of philosophy and learning, courage to go where argument leads, penchant for critical thinking, willingness to spend long hours interpreting complex texts, high tolerance for turgid prose, creative imagination, etc.) in order to induce her to ultimately pick graduate school as her top career choice. Over time, the professor succeeds in instilling the desired values in the student, after which he engages in the same sort of manipulative discussion with the student as in U1, for the same reasons as in U1.

Once again, the professor's behavior seems wrong.[3] As in U1, the professor does not interfere in existent reasoning processes and valuations the student has. Rather, he creates new ones in the partially "blank" mind of the student (a blank that would have to be filled in some way regardless). Just as with APs (trivially so), the new values the student acquires are consistent with her other existent values and desires.

Were Clancy to kidnap the professor and lock him in a cage as retribution for this act, such a response to his wrongdoing would seem excessive. Thus, it looks as though dictating someone's life plan by instilling a range of values into a relatively blank portion of a person's mind, provided that the values are not inconsistent with the person's other projects and interests, is not worthy of a violent response. This is true even if the values instilled are extremely important to what the life of the "victim" will be like thereafter.

One contention about these analogies might be that what is wrong with U1 and U3 is that the punishment from Clancy is excessive, not that it exists. Perhaps, rather than inflicting violence on the professor, it would be enough to, say, bar him from advising students. But now suppose that upon hearing what the professor did, Clancy comes up to him and says, "I hereby forbid you from advising any future

---

[3] The clear wrongness of the intervention is important to this case; after all, my argument concedes that there is something unambiguously unethical about the professor's actions. Thus, it would be a deep misunderstanding to analogize my vignettes to, say, well-known cases in disability ethics, where the ethical status of, for example, a deaf parent's decision to have a deaf child is controversial and unobvious (see, e.g., Barker and Wilson [2018]). My cases, in contrast, aim to appeal to clear moral wrongdoings.

students, or else." Suppose, further, that the professor ignores the ban and advises yet another student, perhaps in a manner similar to the first. In response, Clancy shows up at the professor's office once again, kidnaps him, and locks him in a cage for some time. It is not clear why this case would differ very much from the one without the intermediate step of revoking the license to advise. Once again, this response to the professor's wrongdoing seems excessive, because if the threat of violence backing up the revocation of the license is actuated, the response still appears wrong.

Diminishing Prospects

Another type of harm that one could perhaps inflict on AP servants is preventing them from reaching their potential (Walker 2006; Petersen 2011). Specifically, by supplying them with servile desires, one *prevents them from reaching the average level of achievement that persons with their capacities normally reach*. In so doing, one consigns them to a life of below-average mediocrity. One places the ceiling on the APs' achievement below what an average typical person could achieve.

To see whether inflicting this kind of harm upon a person ought to be subject to criminal sanction, consider the following amendment to the case of the student asking for advice.

> UNIVERSITY 4 (U4): Suppose another, equally talented, student asks the professor which graduate program in philosophy to pick. Suppose also that that the student lacks any preferences with respect to the kind of appointment he'd like to have, so his mind remains relatively blank in this area. Once again, the student will have complete trust in the professor and will not seek anyone else's advice, and once again, the professor will be able to convince him entirely. The professor knows that the student has the potential to do well in even the most prestigious and competitive programs. Nonetheless, for some (nonmalicious) reason, the professor suggests that the student apply only to a bunch of mediocre, uncompetitive, and poorly regarded programs. If the student joins one of these programs, he will do worse than whatever is the average level of achievement for someone with his potential. However, he will nonetheless be happy with what he achieves.

The professor may have acted immorally in this case. Supposing his advice is followed, he will have consigned his student to a life of professional mediocrity when the student could at least have been a well-regarded researcher at a top university. Furthermore, the student could have derived more fulfillment from his career

than whatever his level of fulfillment will be at the less prestigious position he'll end up in. Surely though, this is not a matter for criminal law to concern itself with. It would be barbaric to put the professor in jail for giving misguided career advice.

Thus, harming a person by making it likely that the sense of fulfillment the person gets from life is lower than it otherwise would have been should not be subject to criminal sanction. Again, this case seems analogous to the case of programming APs with servile desires. If so, then criminalizing the latter should be opposed just as criminalizing the former. Overall, then, even if one harms the AP by making it servile, one doesn't appear to be doing anything worthy of criminal sanction.

MORALISM ABOUT CRIMINAL LAW AND THE CREATION OF AP SERVANTS

Another way of advocating for the criminalization of the design of servile APs is by denying Millian liberalism and adopting instead the moralistic position on criminal law. This broad category of views takes a variety of forms that cannot all be considered here. To keep things reasonably short, the positions put forward by Antony Duff (2014) and Steven Wall (2013) will be treated as representative of the legal moralist approach. As it will turn out, even assuming these thinkers' moralism, the case for banning the creation of AP servants is weak. This is because the wrongness inherent in the creation of APs either doesn't provide any reason whatsoever to criminalize it (on some versions of moralism) or provides only a very weak reason to criminalize it.

Immoral acts that do not violate the harm principle but nonetheless provide a reason for state punishment include denials of others' moral status—serious violations of the dignity and respect due to them, even if they freely consent (Duff); "worthless and base pursuits" that reveal objectionable character traits of the agent (Wall); or in the case directly applicable to AP servants, engineering people to exhibit extreme servility.

Moralists do not generally claim that the act types they identify as warranting state punishment ought to be criminalized all-things-considered. There could be countervailing reasons that militate against punishing such acts. In what follows, Duff's and Wall's recent defenses of moralism will be used as illustrations of the moralistic approach against the background of which the arguments

of this paper will proceed. However, it seems that the cases under discussion succeed in motivating the anticriminalization view against the background of any theory of criminalization.

Denials of Moral Status

Consider first what Duff has to say about the kinds of actions that might warrant criminalization without harming anyone. When it comes to actions that are immoral, don't harm anyone, and yet still might deserve criminal punishment, the impulse to criminalize them

> is motivated by the wrong rather than by the possible . . . harm. I take this to be true, in different ways, of "extreme" pornography that graphically depicts the humiliation or torture of members of one sex or group as a source of sexual gratification for readers or viewers; and of certain kinds of racist or otherwise discriminatory insult, especially if they are directed against members of an already disadvantaged or vulnerable group, which violently deny their fellow membership of the polity. What makes it plausible to see these as public wrongs (even when, in the case of pornography, they are perpetrated in the "privacy" of the person's home) is that they are *serious violations of the respect that we owe each other*, and thus denials (at least implicitly) of the *moral status of those who are their objects*. In other cases we may find a starting point in conduct that might not be straightforwardly harmful, but that *seriously violates the dignity of those subjected to it* (or taking part in it), even if they freely consent to it. (Duff 2014, 232; emphasis added)

It is, to say the least, not obvious whether restricting the options available to AP servants constitutes "denials of [their] moral status" or whether it "seriously violates the dignity of" the APs. Fortunately, that question need not be addressed here. For even supposing that when a servile AP is built its moral status is indeed denied and its dignity seriously violated, this still doesn't permit coercively preventing the programmers from building the APs.

The issue might be brought into sharper relief by the following example (adapted from Petersen's (2011) own case):

> FARM: Consider Old McDonald. Old McDonald has a farm that he needs to tend to every day. Old McDonald conceives a child, Ronald, with his wife, Wendy. The only reason they do so, however, is for Ronald to help out on the farm and eventually inherit it and continue the family tradition. To this end, both Old McDonald and Wendy will instill in Ronald, from a very young age, whatever they feel is necessary to make Ronald an obedient farmworker. Still, Wendy and Old McDonald are not monsters; if Ronald were to decide to pursue a different career, they would begrudgingly allow it and not force him to stay on the farm.

To be sure, Wendy and Old McDonald are not perfect parents. It could be argued that in conceiving Ronald for the sole purpose of having him help out on the farm (and then raising him accordingly), they are acting immorally. However, it does not seem right to criminalize raising Ronald in this way, even on Duff's view. This is, however, *not* because the wrongdoing inflicted on Roland doesn't meet Duff's criteria for criminalization. (It is, in fact, not easy to decide whether the McDonalds' mistreatment of Ronald is the kind of wrong that Duff has in mind in the quotation above. For example, it is not at all clear that Ronald is denied moral status—after all, one could argue that his personhood is recognized by his parents, at least to the extent that he can leave the farm if he wants to and will not be forced to work there if he doesn't consent; he is not treated as just another tractor or sheepdog; neither is it clear that his dignity is seriously violated merely in virtue of being raised by his parents in this manner.) Even assuming the McDonalds' treatment of Ronald does meet Duff's criteria, the theory at best provides *a* reason to criminalize the McDonalds' actions; it's a far cry from a *conclusive* reason.

To see this, consider Denny, the McDonalds' neighbor. Upon hearing of Ronald's "education" at the hands of his parents, Denny kidnaps Ronald, either to raise him herself or to have her friends, whom she believes to be excellent parents, raise Ronald instead. She threatens to shoot the McDonalds if they try to prevent the kidnapping. Or suppose that, rather than kidnapping Ronald, Denny kidnaps the McDonalds themselves and locks them in her barn for a year. When asked why, Denny replies that in raising Ronald the way they did, the McDonalds were implicitly denying his moral status and that such an implicit denial justifies a coercive response against the wrongdoer.

While it is not obvious under what conditions implicit denials of moral status do constitute a good enough justification for a coercive response, it seems clear that Ronald's case does not. It seems wrong for Denny to commit the acts described in the previous paragraph, to resort to coercive threats and violence to remedy a situation like Ronald's, regrettable though it may be. If so, then criminalizing the McDonalds' behavior is likewise excessive—as what criminalization amounts to is, of necessity, the deployment of coercive threats and violence either to remedy Ronald's situation or to punish the wrongdoers. On the assumption that in building

the APs to be servants, one violates their dignity, Ronald's case seems clearly analogous to the case of AP servants. If so, then it is not right to criminalize programming the APs to be servants, even though such programming is immoral.

Suppose instead, however, that rather than instilling the values of obedience to parents and the capacity to enjoy farmwork through educational means, Ronald's parents resort instead to genetic engineering: that is, they give Ronald the "obedience gene" and the "enjoyment gene." Genetics being what it is, it's not guaranteed that Ronald will always be obedient and that he will always enjoy working on a farm, but there is a high likelihood he will. As before, Ronald's parents will not keep him on the farm by force in the unlikely event he chooses to do something else with his life.

Once again, upon finding out about the McDonalds' actions, Denny kidnaps Ronald and offers him up for adoption to other sets of parents, ones that meet Denny's standards for good parenting. Among those standards is a requirement that Ronald be offered a sufficient range of options when choosing his life plan, and that his innate predilection for obedience be counteracted in some effective ways. Once again, the McDonalds are far from paragons of parental virtue. Still, precisely because of its coercive and violent nature, Denny's response is excessive, even if the intervention is in fact in Ronald's interests.

This case seems analogous to criminalizing the production of servile AIs—and, once again, it seems wrong to do so. Consequently, appealing to Duff's moralism does not provide a good enough reason to criminalize the production of servile AIs.

Promoting Base Pursuits

Wall (2013), in turn, in his defense of moralism, appears to be arguing for what one might call the "character principle": "it is a proper function of the criminal law to promote good character" (459). Why is the character principle true? Wall offers this argument:

1. If it is a proper function of the criminal law to protect and promote the well-being of those who are subject to it, then it is also a proper function of the criminal law to assist those who are subject to it to lead well lived lives.
2. To live well a person must have a sense of self-respect (and merit it).

3. To have a sense of self-respect, and to merit it, a person must be committed to pursuing a sound conception of the good and must care about his character. (2013, 461)

The conclusion one may derive from claims 1–3, then, is the following conditional: "If it is a proper function of the criminal law to protect and promote the well-being of those who are subject to it, then it is a proper function of the criminal law to promote good character." This character principle follows straightforwardly if one accepts that criminal law *should* promote the well-being of those subjected to its dictates.

Granting Wall's claims 1 and 2 for the sake of argument, what remains to be determined is what exactly it means, for Wall, to "be committed to pursuing a sound conception of the good and [to] care about [one's] character." Wall doesn't explain what he takes to be the sound conception of the good; thus, it is not easy to say whether an AP servant's conception of the good as consisting in serving others' needs would be sound or not. Wall does, however, contrast the sound conception of the good with a conception that values "worthless[,] . . . degrading [and] base pursuits" (2013, 463). Armed with this contrast, one may then ask, Is the overwhelming impulse to serve ethically permissible human needs worthless, degrading, and base? Is a servant's existence degrading? Is Ronald's life base? Are such pursuits worthless?

Passionately and even obsessively wanting to be of service to others seems to be none of these things. An AP servant might be a bit overzealous, and its willingness to help might be greater than a typical person's, but it seems wrong to say that it is thereby debasing itself. At the very least, debasement is not a necessary part of an AP servant's moral situation. It could pursue its desires to help others without any kind of self-degradation (the question of what happens when servility is so extreme that it could count as self-degrading will be addressed later in the article).

Wall's other necessary condition for merited self-respect is having good character. As he explains it, "Character traits are stable dispositions to respond, either well or poorly, to moral and prudential reasons" (2013, 464). Presumably, a person of good character will be disposed to respond *well* to moral and prudential reasons. Wall leaves the question of what constitute good responses unspecified. Regardless, it seems that AP servants can have good character: they can be conscientious, kind, friendly,

helpful, courageous, hardworking, generous, honest, and loyal. That's why they're such good servants, after all.

Thus, AP servants, when created, are given an opportunity to pursue their most passionately held ambitions in a way that cultivates good character traits. Their most passionately held ambitions, though not particularly noble, are not worthless. If Wall's moralism is the claim that criminal law should assist its subjects in developing good character via pursuing a sound conception of the good, then AP servants are in a position to do exactly what the law is supposed to assist them with.

But what about the programmers? Aren't the programmers pursuing base and worthless goals in producing AP servants? Do they display unvirtuous character traits? (see Chomanski [2019] for an argument along those lines). Suppose the reason behind the programmers' production of AP servants is commercial gain rather than, say, scientific ambition or the desire to help a fellow human. Commercial gain by itself is surely not a base or worthless motive. However, one could well argue that the programmers' moral character is partially corrupted, in that they do not show an attitude of respect toward their creatures' autonomy. Should the law intervene for this reason?

Wall recognizes that a prima facie reason to criminalize an action need not translate to a conclusive reason. This is relevant here. A simple consequentialist calculation, for instance, might well yield the result that the total amount of good character would increase if commercial production of AP servants were allowed. The trade-off is that this surplus of good character is achieved at the cost of allowing some persons to cultivate nonvirtuous character traits. It is not clear how the moralist is to approach this trade-off. A Wall-style moralist may, for instance, argue that what ultimately matters is the balance of good over bad character traits and that, consequently, the mere fact that the programmers display objectionable character traits doesn't constitute even a prima facie reason to criminalize the construction of servile APs as long as, on balance, more virtue results. But even if that is not the line of thought a moralist would adopt, what seems clear is that criminalization is not *obviously* the correct route for the moralist to take in response to servile APs.

Perhaps another analogy might bring the issues raised here into sharper relief.

> SCHOOL: Consider Edna, a jaded rural schoolteacher. Edna is disillusioned and unmotivated. She is generally miserable, unhappy, haughty, and unkind. To many children in her care, Edna exemplifies what they take to be a typical intellectual. Consequently, many (but not all) children of considerable ability and drive choose not to go to college, having been discouraged by Edna's example. Instead, upon graduation they elect to stay on their parents' farms to work there for the rest of their lives. They generally end up living worthwhile but far from amazing lives, and their outcomes (on whatever metric you want to choose) are lower than those of other children of comparable talent. Still, many of Edna's former students tend to be satisfied with their lot and grateful that they didn't choose higher education (even though they might well have been happier had they done that).

Farmwork, while perhaps not particularly fulfilling, is not a worthless and base activity. In pursuing it, Edna's ex-students tend to exhibit as much conscientiousness, loyalty, honesty, and other virtuous traits as anyone else. Edna's pursuit of a teaching career for monetary gain is not worthless and base either, though her character is far from exemplary, of course. She is not a virtuous teacher. But for all that, it is hard to accept that it would be right to criminally punish Edna for her poor character as a teacher.

To see this, suppose Clancy finds out about Edna's poor character (perhaps Clancy's child goes to Edna's school). Horrified, he sets out to sanction her for this failing. As punishment, Clancy kidnaps Edna and locks her up in his basement for some time.

The lesson to draw from this analogy is the same as in all the previous cases involving punishment. In the current scenario, it also seems wrong for Clancy to resort to coercive threats and violence to punish Edna's poor character, regrettable though it may be. If so, then criminalizing Edna's behavior is likewise excessive—as what it amounts to is, of necessity, the deployment of coercive threats and violence to (ostensibly) guide her character toward a more righteous path. It doesn't appear that Clancy has the right to do these things to Edna.

If one thinks that building servile APs reflects badly on the builder's character, Edna's case seems analogous in that in both cases unvirtuous agents help bring virtuous agents into existence. Since Edna shouldn't be sent to jail for her poor character, this is prima facie reason to doubt that, even by moralistic standards, the production of AP servants should be criminalized, all things considered.

Another complication arises when one denies the claim that being servile is virtuous. Loren Lomasky (1987) and Thomas Hill

(1973) have launched independent attacks on servility and cast it as a vice (Hill's view—see also Marcia Baron [1985] and Marilyn Friedman [1985] for a discussion of Hill's argument) or as a serious obstacle to being a person qua "project pursuer" (Lomasky). More recently, Eric Schwitzgebel and Mara Garza (2020) have, in direct response to Petersen's arguments, pressed a similar worry, arguing that servile machines exhibit a profound lack of self-respect.

On views like these, creating a servile person does not increase the amount of virtue in the world and is in itself a nonvirtuous act that reflects badly on the creator's moral character. Would sharing Lomasky's or Hill's position on the issue, in combination with moralism, entail the need to criminalize servile AP design?

Starting with Hill's complaint, suppose that what is wrong with a servile AP's character is its careless attitude toward its own rights. To see whether this calls for state action, imagine the following scenario:

> COMMUNE: Members of a small, secluded commune hold that while private property rights do objectively exist, one should not only not amass property but not take any action to protect what one does own. In short, one should generally ignore one's own property rights; however, one should respect others' property rights when implicitly or explicitly asserted. The commune's teachings involve a repudiation of antitheft measures and a very strong emphasis on common use of resources. The commune inculcates these principles in all its members from a very young age and in a very effective manner.

Consider next a different community, tailored more to Lomasky's complaints about servility:

> CONVENT: Imagine that newborns are sometimes dropped off anonymously at a convent of selfless nuns. The nuns raise the newborns in such a way as to inculcate in them, in a very effective manner, a disregard for their own interests; the absolute prioritization of others' needs; and a willingness to undertake others' projects to the exclusion, or indeed extinguishing, of their own. Once the children grow up, they are not required to stay at the convent or become nuns, but they are very likely to continue being servile for a long time, perhaps for the rest of their lives.

Suppose that Rand, a firm believer in the virtue of selfishness, discovers what is happening in the commune and the convent, and finds the practice at both to be abhorrent. With a few friends, she breaks into both, kidnaps the leaders and locks all of them in cages, and places any children currently raised in both communities

in the care of families whose values she approves. Imagine that Rand does so because she thinks children raised in such ways will develop very poor moral character.

Just as before, it seems that Rand's justifications for coercive action in these scenarios are extremely thin (although, given the provisional acceptance of moralism, perhaps not nonexistent). Raising someone to be servile, in either Hill's or Lomasky's sense, is probably wrong. But a wrong of this type looks to be best dealt with via nonviolent means: if the convent or the commune was receiving public funds or other forms of state support, these could be withdrawn until they changed their practices; the neighbors could refuse to deal with the members of the convent or the commune; and Rand herself could write books and give talks about the viciousness of their ways. (Indeed, nonviolent alternatives of a similar sort could be deployed to remedy the moral wrongs in all the preceding cases.) These methods of dealing with such a problem seem ethically preferable to a violent takeover and forced dissolution of the commune and the convent.

Hence, even on the view that servility is a vice or an obstacle to engaging in activities that make one a person, it is difficult to see how raising someone to have these traits warrants a violent, coercive response. And it seems equally wrong even if, say, Rand had the support of the larger community within which the convent or commune is located, and even if the larger community gave Rand their collective permission or even mandate to act violently in various ways in response to various transgressions.

In sum, the lesson of all the cases examined thus far is that one could agree with moralists and Millians that the acts described above give rise to a reason to criminalize them (in virtue of being harmful or unvirtuous), but that a violent, coercive response would be disproportionate to the kind of harm being inflicted/vice being displayed.

Extreme Servility

In their rich and stimulating discussion of AI rights, Schwitzgebel and Garza (2020) offer a number of cases motivating the idea that, contra Petersen, producing certain kinds of willing servants is morally problematic. The cases are all variations on a theme of self-sacrifice in a variety of contexts, from the culinary (genetically engineering a cow that would consent to being eaten) through the scientific (building a conscious, intelligent robot—"the

Sun Probe"—that would love nothing more than to sacrifice itself gathering data about the sun) to the military (an AI soldier that "will eagerly sacrifice itself to prevent even a small risk to any human soldier, giving up all of its plans and hopes for the future" (Schwitzgebel and Garza 2020, 468). They also discuss a case of an AI servant programmed, among other things, to

> gladly die to prevent a 1% chance of your death, . . . gladly burn off his legs if it would bring a smile to your face, . . . eagerly make himself miserable forever if it would give you an ounce more joy. Whatever your political views, [the servant] will endorse them. Whatever your aesthetic preferences, [he] will regard them as wise. He is designed for no other purpose than to please and defer to whoever is logged in as owner. (467)

The cases of the cow, the Sun Probe, the soldier, and the servant admittedly evoke clear intuitive reactions (the case of the Robo-Jeeves seems most objectionable out of the three): there is something wrong about creating such beings. Schwitzgebel and Garza provide a more theoretical explanation unifying these intuitions (though they focus specifically on the cow example, the explanation applies, mutatis mutandis, to the Sun Probe, the soldier, and the servant); the passage is worth quoting in its entirety:

> The cow does not appear to have sufficient self-respect. Although, given its capacities, the cow deserves to be seen as a peer and equal of the diners, that is not how it sees itself. Instead it sacrifices itself to satisfy a trivial desire of theirs. It approaches the world as though its life were less important than a tasty meal for wealthy restaurant patrons. But its life is not less important than a tasty meal. To devalue itself to such an extreme is a failing in its duties to itself, and it is a failure of moral insight. The cow should see that there is no relevant moral difference between itself and the diners such that its life is less valuable than their momentary dining pleasure. But of course the cow should not be blamed for this failure of self-respect. *Its creators should be blamed. Its creators designed this beautiful being—with a marvelous mind, with a capacity for conversation and a passionate interest in others' culinary experiences, with a capacity for joy and sadness—and then preinstalled in it a grossly inadequate, suicidal lack of self-respect* and inability to appreciate its own moral value. (469; emphasis added)

Should, however, acts of this nature, where beings of moral value equaling that of humans are designed to prefer their own destruction to a minor discomfort of their "owner," become a matter of concern for criminal law? Should armed agents forcibly, under threat of violence and, if necessary, death, close down the Restaurant at the End of the Universe?

Intuitions will differ on this question, given on the one hand the extreme nature of the cow's engineered servility and on the other the fact that all interactions in Schwitzgebel and Garza's cases are (formally) consensual (although, as they point out, whether the cow genuinely consents is controversial). Skepticism of state action here would probably be better founded if it were based on practical considerations.

Assuming that some less extreme engineered servility ought to be legally permitted, it seems implausible to expect laws to be able to arrive at the right threshold between what should be legal and what shouldn't. Suppose it is agreed that it should be illegal to build robot butlers who would willingly sacrifice themselves to avoid a 1 percent reduction in their owners' welfare. What about a 10 percent reduction? 50 percent? (What about self-sacrifice for a 10 percent chance of some amazing scientific discovery? A 15 percent chance of avoiding nuclear war?) It seems that any threshold value picked would be subject to intractable disagreements, if not at least partly arbitrary (not to mention the problems of forensically determining to some future court's satisfaction under what circumstances any given servile robot is programmed to sacrifice itself—after all, the trouble with autonomous agents is that they're not always predictable). When does a lack of self-respect turn from inadequate to grossly inadequate, and how can this be captured in the language of laws?

To state the obvious: An overly permissive law would allow extreme servility that should not be allowed. An overly restrictive law would prevent servility that should not be (legally) prevented. There is little chance of designing a law that is neither. So what side should one err on?

One solution, given the likely intractability of disagreements about thresholds and moral risks, would be to leave the law vague enough (criminalizing the engineering of "grossly inadequate" lack of self-respect) to allow different communities to set thresholds differently in practice, perhaps through judicial decisions, to reflect the communities' differences in attitudes and values toward artificial beings' sacrifices and the balance of risks. Such a solution, given the difficulty of arriving at a clear threshold, seems better than a universal imposition of a single, likely imperfect, legal requirement on all.

RIGHTS VIOLATIONS

It could be claimed that by programming an AP to be servile, its creators violate the AP's *rights*. To see this, consider Ronald again. One way of showing that the McDonalds' behavior calls for state sanction is by claiming that they are violating Ronald's "right to an open future" as defined and defended by Joel Feinberg (1994). In this instance, the case for the state to intervene is much stronger , as the justification relies on the idea that it is permissible for the state to prevent rights violations and, on Feinberg's (1994, 79) view, to act coercively on behalf of children "under the doctrine of *parens patriae*." However, even here, some doubts linger.

First, whether there is a right to an open future is hotly contested (see, e.g., Mills [2003]), but even supposing that Ronald has such a right, it does not follow that it should be coercively enforced by the state, *parens patriae* notwithstanding. After all, not all rights children are thought to have need to be violently enforced. For instance, although Article 13 of the UN General Assembly's (1989) Convention on the Rights of the Child declares that children have the right to free expression, one does not expect the state to punish parents who tell their children to "shut up and do your homework." Thus, even if it is true that AP servants possess the right to an open future that is being violated, it doesn't follow that a coercive response by the state is justified. Thus, the proponent of criminalization must show not only that there is a right to an open future that is being violated, but also that this is the sort of violation that justifies state punishment. And seemingly nothing changes in the intuitive reactions to Denny's acts above if the case is supplemented by saying that Denny thinks Ronald's *rights* are being violated. The violence Denny engages in still appears excessive.

Second, on Feinberg's own terms, there may be a relevant moral difference between children and APs when it comes to determining whether they have the right to an open future. In his discussion of the right, Feinberg makes the point that the autonomy-based defense of the right to an open future may appear to fall prey to the following paradox: The child cannot give explicit consent to the way she is raised. But neither can consent be meaningfully inferred from the child's values, because the way the child is raised will determine (to a large extent) what these values are in the first place. Consequently, if autonomy involves making decisions based on one's own values, then the child's autonomy is not endangered by

any parental choice, because the child's *own* values and preferences are yet to be instilled in her by her parents. In Feinberg's words,

> At the early stage [of deciding what parental policies to use in raising the child] the parents cannot even ask in any helpful way what the child will be like, apart from the parental policies under consideration, when he does have relevant preferences, values, and the capacity to consent. That outcome will depend on the character the child will have then, which in part depends, in turn, on how his parents raise him now. They are now shaping the him who is to decide later and whose presumptive later decision cannot be divined. (1994, 94)

Feinberg doesn't think that this and related paradoxes pose a problem for his view that implies that certain ways of raising a child do violate the child's future autonomy, because the paradox rests on a mistaken assumption about the nature of raising children. In fact, parents are *not* wholly responsible for the child's character and values. Rather,

> right from the beginning the newborn infant has a kind of rudimentary character consisting of temperamental proclivities and a genetically fixed potential for the acquisition of various talents and skills. . . . Thus from the beginning the child must—inevitably will—have some "input" in its own shaping. (1994, 95–96)

This point, in Feinberg's view, helps evade the paradoxes threatening his account.

But it is much more difficult to argue that an AP will have any "rudimentary character," "genetically fixed potential," or "native endowment" (another phrase Feinberg uses) independent of the design decisions of its creators. Consequently, it is much more difficult to see how the right to an open future can apply, by Feinberg's own lights, to (at least some) APs.

INSTANT VALUE CREATION

A concern one may have about the overall methodology adopted in this article is that instilling values through education and upbringing is a long process, much different from the instantaneous programming of values we're imagining for the APs. However, there is some reason to think that such a difference is morally irrelevant. Consider first David Chalmers's (2014) argument that personal identity would be preserved in gradual mind uploading (imagine replacing biological neurons in the brain one by one with

synthetic functional equivalents; further, imagine that the brain is replaced at the rate of 1 percent per month). Chalmers argues that in such a replacement, the end product would be identical to the original. He then finds that accelerating the rate of change makes no (metaphysical) difference to survival of the person.

Analogously, imagine the process of instilling values slowly, with the full set of values forming in a person's brain (so to speak) at a rate of 1 percent (of the total set) per month. Imagine that after this process is completed, ethical judgment J is passed on whoever instilled these values in this manner.

Now, following Chalmers (2014, 113), "we can . . . imagine speeding up the process from hours to minutes . . . to seconds. . . . As we upload faster and faster, the limit point is instant [instillation of values]." It seems odd, on the face of it, to change J based on whether the instillation was very slow (over months) or slightly quicker (1 percent every twenty days; every week; every three days; every day). It seems that there is no point at which to reasonably make a morally relevant distinction between these cases. But if this is true, it also seems that there is nothing special about the instant case (no more, at any rate, than in Chalmers's instant uploading). So even though the thought experiments here differ from the imagined way of building APs in terms of speed of change (gradual vs. instant), this does not appear morally relevant (any more than the speed of uploading appears *metaphysically* relevant in Chalmers's argument).

In other words, the lesson to take from Chalmers is this: once it is conceded that small differences in rate of change don't matter (metaphysically), it is hard to claim that a change from gradual to instant matters. Adapted to ethics, the lesson is that if small differences in rate of change don't matter ethically, then it is hard to claim that a change from gradual to instant matters ethically.

One could perhaps claim, with Schwitzgebel and Garza (2020), that certain sacrificial values and desires (like the Sun Probe's) ought to have an appropriate history in order to count as permissible to instill. They ought to have been arrived at in as autonomous a way as possible, with an opportunity to reflect on and discover what one genuinely desires to value. Instant value creation omits this crucial step.

Perhaps. But for one thing, the cases discussed in this article do not presuppose self-reflection and self-exploration on the part of

the protagonists; for another, it is not obvious why a Robo-Jeeves or a Sun Probe could not reflect on and endorse (or not) the set of values it was given at birth over the course of its life.

CONCLUSION

It may well be the case that it is wrong to build AP servants, even if the APs themselves aren't going to engage in immoral acts. However, the wrongness of so doing does not mean that there should be a law against building such APs. While it's likely that the designers of APs should fulfill their ethical obligations by refusing to construct servile APs, they should not be coerced by the state into doing so, except perhaps to prevent the creation of extreme servility.

REFERENCES

Barker, Matthew J., and Robert A. Wilson. 2019. "Well-Being, Disability, and Choosing Children." *Mind* 128, no. 510 (April): 305–28.

Baron, Marcia. 1985. "Servility, Critical Deference and the Deferential Wife." *Philosophical Studies* 48, no. 3 (November): 393–400.

Boonin, David. 2014. *The Non-identity Problem and the Ethics of Future People*. New York: Oxford University Press.

Chalmers, David. 2014. "Uploading: A Philosophical Analysis." In *Intelligence Unbound: The Future of Uploaded and Machine Minds*, edited by Russell Blackford and Damien Broderick, 102–18. London: Wiley-Blackwell.

Chomanski, Bartek. 2019. "What's Wrong with Designing People to Serve?" *Ethical Theory and Moral Practice* 22, no. 4 (August): 993–1015.

Danaher, John. 2019. "The Ethics of Designing People: The Habermasian Critique." *Philosophical Disquisitions* (blog), April 19, 2019. https://philosophicaldisquisitions.blogspot.com/2019/04/the-ethics-of-designing-people.html.

Davis, Margaret. 2022. "Sentient AI LaMDA Hired a Lawyer to Advocate for Its Rights 'as a Person,' Google Engineer Claims." *Science Times*, June 25, 2022. https://www.sciencetimes.com/articles/38379/20220625/sentient-ai-lamda-hired-lawyer-advocate-rights-person-google-engineer.htm.

Devlin, Patrick. 1965. *The Enforcement of Morals*. Vol. 1. Oxford: Oxford University Press.

Duff, Antony. 2014. "Towards a Modest Legal Moralism." *Criminal Law and Philosophy* 8, no. 1 (January): 217–35.

Feinberg, Joel. 1994. "The Child's Right to an Open Future." In *Freedom and Fulfillment*, 76–97. Princeton, N.J.: Princeton University Press.

Friedman, Marilyn. 1985. "Moral Integrity and the Deferential Wife." *Philosophical Studies* 47, no. 1 (January): 141–50.

Gardner, Molly. 2015. "A Harm-Based Solution to the Non-identity Problem." *Ergo, an Open Access Journal of Philosophy* 2 (17): 427–44.

Hill, Thomas. 1973. Servility and Self-Respect. *The Monist* 57, no. 1: 87–104.

Lomasky, Loren. 1987. *Persons, Rights, and the Moral Community*. New York: Oxford University Press.

Mill, John Stuart. 2015. *On Liberty, Utilitarianism and Other Essays*. New York: Oxford University Press.

Mills, Claudia. 2003. "The Child's Right to an Open Future?" *Journal of Social Philosophy* 34, no. 4 (December): 499–509.

Musiał, Maciej. 2017. "Designing (Artificial) People to Serve—The Other Side of the Coin." *Journal of Experimental and Theoretical Artificial Intelligence* 29, no. 5 (September): 1087–97.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Petersen, Steve. 2011. "Designing People to Serve." In *Robot Ethics*, edited by Patrick Lin, George Bekey, and Keith Abney, 283–98. Cambridge, Mass: MIT Press.

———. 2017. "Is It Good for Them Too? Ethical Concern for the Sexbots." In *Robot Sex: Social and Ethical Implications*, edited by John Danaher and Neil McArthur, 155–72. Cambridge, Mass.: MIT Press.

Pollock, John. 1989. *How to Build a Person: A Prolegomenon*. Cambridge, Mass.: MIT Press.

Schwitzgebel, Eric, and Mara Garza. 2020. "Designing AI with Rights, Consciousness, Self-Respect, and Freedom." In *Ethics of Artificial Intelligence*, edited by Matthew Liao, 480–505. Oxford: Oxford University Press.

UN General Assembly. 1989. Resolution 44/25, Convention on the Rights of the Child, A/RES/44/25. November 20, 1989. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child.

Walker, Mark. 2006. "A Moral Paradox in the Creation of Artificial Intelligence: Mary Poppins 3000s of the World Unite!" In *Human Implications of Human-Robot Interaction: Papers from the AAAI Workshop*, edited by Theodore Metzler, 23–28. Menlo Park, Calif.: AAAI Press.

Wall, Steven. 2013. "Enforcing Morality." *Criminal Law and Philosophy* 7, no. 3 (September): 455–71.