AI & SOCIETY
https://doi.org/10.1007/s00146-022-01575-z

OPEN FORUM

Can we design artificial persons without being manipulative?

Maciej Musiał[1]

Received: 25 April 2022 / Accepted: 26 September 2022

Abstract
If we could build artificial persons (APs) with a moral status comparable to this of a typical human being, how should we design those APs in the right way? This question has been addressed mainly in terms of designing APs devoted to being servants (AP servants) and debated in reference to their autonomy and the harm they might experience. Recently, it has been argued that even if developing AP servants would neither deprive them of autonomy nor cause any net harm, then developing such entities would still be unethical due to the manipulative attitude of their designers. I make two contributions to this discussion. First, I claim that the argument about manipulative attitude significantly shifts the perspective of the whole discussion on APs and that it refers to a much wider range of types of APs than has been acknowledged. Second, I investigate the possibilities of developing APs without a manipulative attitude. I proceed in the following manner: (1) I examine the argument about manipulativeness; (2) show the important novelty it brings to a discussion about APs; (3) analyze how the argument can be extrapolated to designing other kinds of Aps; and (4) discuss cases in which APs can be designed without manipulativeness.

Keywords AI ethics · Artificial persons · Designing · Manipulativeness

1 Introduction

In the movie "Bicentennial Man" (Columbus 1999) a robot named Andrew, played by Robin Williams, becomes recognized as morally equal to human beings. Andrew possesses human-level intelligence and experiences sensations and emotions. As a matter of fact, recognizing him as a human person confirms that although he is an artificial being, he does not differ from a natural human being in any way that is relevant in terms of ascribing moral personhood. It is extremely unclear whether such robots with human-level intelligence and other properties that will enable them to be recognized as persons morally equal to humans will ever actually exist. Yet, some people—not only in sci-fi and pop culture, but also in academic fields (Schwitzgebel and Garza 2015; Tavani 2018; Gellers 2021; Gordon and Gunkel 2021; Gordon 2022; Mamak 2022)—are discussing ethical dilemmas that would arise if such beings could be developed. In general, many of these considerations revolve around the question, how should we design such artificial persons in an ethically right way, considering not only the well-being of humans and the environment, but also of these artificial persons themselves? In particular, some wonder whether we should design them as our servants.

The scenario and the questions mentioned above are the key issues of the debate to which this paper would like to contribute. This debate is clearly different from the majority of academic debates about robots. The debates that refer to robots who are persons tend to ask what properties such entities should possess to achieve such—or similar—moral status and in what way their personhood could and should be recognized, but the debate this study contributes to begins with the assumption that such entities will exist and the questions signalized above will be answered. Moreover, while most other debates focus on how robots and the way they are designed will affect the well-being of humans, the current discussion concentrates—at least for the most part—on the well-being of robotic artificial persons.

Hence, the initial and main question of the debate that this study engages with is: assuming that we could build robots that would possess all necessary features to consider them (artificial) persons (APs), would it be good for these robots to be designed to desire to be servants of human beings (AP servants)? The discussion has been initiated by Mark Walker

✉ Maciej Musiał
m.musial@amu.edu.pl

[1] Adam Mickiewicz University in Poznań, Poznań, Poland

Published online: 17 October 2022

(2006, 2016) and Steve Petersen (2007, 2011, 2017) and continued by John Danaher (2013, 2019), Maciej Musiał (2017), and Sven Nyholm (2020)—the topic has also found some attention outside of typical academic discussions (see: Rini 2017; Bloom and Walker 2018). In a nutshell, Walker and Petersen have provided the two dominant positions in the discussion. Walker argued that designing AP servants is wrong, since it negatively affects their autonomy and harms them, while Petersen disagreed with Walker, claiming that we could design AP servants without harming them and diminishing their autonomy. Danaher, Musiał and Nyholm have generally positioned their views in reference to these two main positions, also focusing on the well-being of AP servants in terms of the process of designing and its potentially harmful and autonomy-diminishing consequences.

Recently, Bartek Chomanski (2019) has provided an interesting contribution to this dispute. In brief, Chomanski argues that (1) there are some plausible conceptions of autonomy according to which artificial persons designed to be servants would not need to lose their autonomy when being designed to desire to serve humans, (2) there are some plausible scenarios in which such a situation would not result in net harm, yet (3) the designers express manipulativeness towards AP servants, and hence (4) the process of designing APs in such a way is unethical. Hence, Chomanski defends a somewhat counterintuitive claim that designers are being manipulative while APs need not be manipulated.

In my considerations, I leave the problems of autonomy, harm and net benefits aside and focus on the issue of the manipulativeness of the designers. My main claim is that Chomanski's argument about designers being manipulative significantly shifts the perspective of the whole discussion about APs and that it refers to a much wider range of kinds of APs than Chomanski acknowledges. Moreover, I examine how to develop APs without being manipulative. To make my point, I (1) examine Chomanski's argument about manipulativeness, (2) show the important novelty it brings to a discussion about APs, (3) analyze how the argument can be extrapolated to designing other kinds of Aps and (4) discuss cases in which APs can be developed without designers being manipulative.

Some may argue that the whole discussion presented here is unrealistic and pointless. Obviously, APs with human-level intelligence who could be recognized as moral persons or possess similar moral status do not exist and it is far from clear whether they will exist in the near future, or at all. It seems that one of the crucial points is that there is no agreement on the possibility of robots with human-level intelligence. Positions on this issue range from radical skepticism and denial of such a possibility to a firm belief about the imminence of a super-intelligent AI that will surpass even the most intelligent human beings. This lack of consensus seems to be clearly reflected in the disagreement among five of the founders of AI—John McCarthy, Marvin Minsky, Oliver Selfridge, Ray Solomonoff, and Trenchard Moore—who were asked about the possibility of human-level AI at the meeting celebrating the 50th anniversary of the conference on which the foundation of AI as a discipline took place (Bringsjord and Naveen 2018). Hence, while the number of scholars who seriously consider a scenario of robots with human-level intelligence and a moral status of a person (or similar) steadily increases (also outside the debate about AP servants, e.g. (Neely 2014; Schwitzgebel and Garza 2015; Gunkel 2018; Gordon 2020, 2021; Gunkel and Wales 2021)), the prospect of such entities remains extremely uncertain. In this case, is there any point in speculating about designing them in an ethical way? I believe there is, regardless of whether such entities will ever exist or not.

Generally, while I do not share Petersen's optimism that APs with human-level intelligence will arrive rather sooner than later, I agree with him that since we cannot rule out with certainty the possibility that APs will exist someday, it is better to be prepared for them than not (Petersen 2007, 44). I also assert that Jacob Turner might be right when he claims that if we had known the dangers of global warming at the dawn of the Industrial Revolution, we might have been able to avoid it (Turner 2019, 35). While such a point of view might seem both excessively cautious (by assuming we should always worry about risks and dangers) and optimistic (by assuming that by knowing the risk we could avoid it), it does not mean it is obviously wrong.

But even if there will never be any APs, and human-level intelligence will never emerge, the discussion about designing them is still not pointless. First, it is connected with some other ethical discussions and may contribute to them. Petersen argues that there is a strong connection between the discussion on APs and some of the current problems facing population ethics (Petersen 2007, 44), while Musiał shows that many problems with designing APs are analogous to those associated with the prenatal enhancement of human beings (Musiał 2017). Second, discussing problems with designing APs enables us to consider whether we would like to be able to develop such APs at all. What I mean is that highlighting the ethical problems with designing robots that are persons and possess human-level intelligence can allow us to make a more informed decision on whether making attempts to develop such APs is worthwhile at all. Hence, speculations such as those presented here may prevent us from spending a lot of resources on creating a world with ethically problematic APs that cause more trouble than they are worth, or may alternatively motivate us to achieve—let's say—a reality in which we are gods who create a new artificial species and make the world more beautifully diversified and less anthropocentric. By imagining possible futures, we can better decide which future we currently want to work towards.

2 Manipulativeness

The main novelty and central claim of Chomanski's contribution is the argument that designing artificial servants is wrong because it involves manipulativeness on the side of designers. In his understanding of manipulativeness, Chomanski follows Marcia Baron's interpretation of Aristotle, which makes the following claim:

"What one gets wrong, in [certain] forms of manipulativeness, and what the corresponding virtuous person gets right, is how much to steer others—and which others, and how, and when, and toward what ends; and more generally, to what extent—and how and when and to whom and for what sorts of ends—to seek to influence others' conduct. … The manipulative person is too ready to think appropriate—or appropriate for him—to orchestrate things so as to lead others to act as he wants them to." (Baron 2003, 48).

In a very clear manner, Chomanski shows that designers of AP servants are being manipulative in the abovementioned sense:

"By programming AIs with overwhelming desires to be servants, their programmers are set on making it close to impossibly psychologically difficult for the AIs to act in any way other than to pursue the life-plan that they are given: even a determined AI servant would likely not pick a different career path. This might be so even if they are programmed with the capacity to reflect on their own desires. Upon being told to build an AI servant, the programmers cannot help but suppose that the servant's decision about what life plane to pursue are for the programmers to make.

The programmers position themselves so as to be able to steer the choices of AI servants with a high degree of efficacy, by constructing the AIs' psychology in the way just specified. They orchestrate the AIs' choice space in such a way that only a very narrow range of options is realistically available to them. The AIs' number of options, when it comes to career choice, is reduced to basically one. The AIs' own capacity to respond to reasons, or to engage in reasoning, is ignored. The act of thus limiting another's options and disregarding their rational capacities is expressive of manipulativeness on the programmers' part—that is, of being willing to orchestrate another's choices in an excessive manner. It is, therefore, wrong" (Chomanski 2019, 1006).

Hence, Chomanski believes that designing artificial servants involves manipulativeness on the side of designers since by selecting specific features of the designee (e.g., the desire to serve), they are "willing to orchestrate another's choices in an excessive manner" by—I take the following features as crucial—"limiting another's options" and "disregarding their rational capacities." Hence, Chomanski contends that designing people to serve is wrong because it involves the manipulativeness of the designers.

A few clarifications on the concept of manipulativeness seem worth making. First, clearly, Baron's account of manipulativeness represents only one of many approaches to this problem. However, I find her approach plausible, and I concede that Chomanski's application of her concept to the problem of designing APs is plausible as well. As a result, I follow their approach here, although I am aware that there are also other paths to approaching manipulativeness. Second, for the sake of clarity, Baron assumes that manipulativeness as a vice is always objectionable (Baron 2003, 39). However, she believes that manipulativeness can sometimes be justified, particularly when applied to avoid significant harm (Baron 2003, 48)—I follow these assumptions as well. Third, both manipulativeness itself, as well as its justification, seem to be matters of degree. Being manipulative is characterized by expressions such as "excessive," "how much," or "too ready," and neither Baron nor Chomanski provides any strict demarcating line between "proper" and "excessive," etc. Hence, manipulativeness can be more or less intensive and probably also more or less objectionable. Analogously, there are diverse ways in which manipulativeness can be justified, and it seems that it can also be justified to a higher or lesser degree. The point here is that while manipulativeness is an objectionable vice, it is neither necessarily a great vice (though it might be) nor an unjustifiable vice (at least in some contexts). Hence, while the presence of manipulativeness in some action—e.g., designing APs—is always objectionable, it should not be simply considered as directly resulting in banning the action in question, particularly because the wrongness of manipulativeness can be somehow justified. Fourth, one could wonder how exactly manipulativeness impacts one's character. I agree with Chomanski when he argues that apart from the fact that manipulativeness is a vice and an expression of a bad attitude towards APs, it can also additionally corrupt one's character (Chomanski 2019, 1010). By expressing a manipulative attitude towards APs, one may generalize and strengthen it, and then express it towards other entities—including animals or humans—as a result. Such a mechanism is widely discussed in the case of relationships with robots in general (e.g. sex robots and their mistreatment), or with some fictional representations in general (playing brutal video games or consuming violent pornography)—a comprehensive analysis of this mechanism has been provided by Danaher (2017a). While this topic has been the subject of an enormous amount of discussion and disagreement, Chomanski seems to be right in his suggestion that if the bad attitude towards a fictional or representational entity raises concerns as to whether this attitude should be extended to human beings or animals, then a bad attitude towards APs—who are recognized as actual persons in the moral sense—raises even more concerns.

Finally, it is worth noticing that the object of manipulativeness in the case of Chomanski's elaborations is desires

(whether manipulativeness would take place in reference to designing some other features of APs, e.g. their embodiment, is a topic for another interesting discussion). Hence, it is worth clarifying how Chomanski (and Petersen) understand the concept of desire. According to both authors, APs' desires would not differ from the desires of human beings except in their origin, since they would not be a result of natural evolution and/or socialization (as they are—at least for now—in the case of human beings), but of intentional "prenatal" designing. Hence, they would be—analogously to human beings, at least in most accounts—mental states that are dispositions for actions. Moreover, fulfilling these desires would result in APs experiencing pleasure. Both Petersen (2017, 290) and Chomanski (2019, 1000) emphasize that these desires could be reflected and, as a result of such reflection, APs could persuade themselves not to follow them, even if—as it is in the case of AP servants—these desires were as strong as the desire for food in the case of human beings. This means that APs desires would have a conscious character, instead of being some unconscious instincts. Hence, to reiterate, the character of the desires designed into APs would not differ significantly from the character of the desires of human beings (except in terms of their origin).

3 Manipulativeness as a shift in the discussion

The claim about manipulativeness is—as it has been already mentioned—an important shift in the discussion. The novelty it brings is not only the argument and its content but also the perspective that differs from the perspective that has been dominating the discussion since its origins. Both Walker and Petersen were focused on the question of whether or not designing artificial servants wrongs them. As Petersen puts it in the title of one of his papers (Petersen 2017), the question is "Is it good for them too?" ("them" refers to AP servants). The same refers to Musiał and Danaher.

Hence, Chomanski's claim about manipulativeness is not focused on the main topic of the discussion to date—the well-being of the designee—and instead focuses on the attitude of the designers. What makes the argument about manipulativeness and the shift in the perspective an interesting contribution is that even if one assumes that APs' autonomy and general well-being are not harmed by designing them as servants, the process of designing AP servants remains objectionable due to the manipulativeness of the designers. This is an important shift, since it shows that even if the consequences of designing AP servants can be ethical both for themselves and for society in general, something can still be wrong with the whole process due to the attitudes of the designers.

Actually, it seems that Chomanski's shift of perspective largely depends on the fact that he—without stating it explicitly—embraces virtue ethics and considers manipulativeness as a vice, following Baron in that regard. Hence, it is worth comparing it with some other accounts that approach similar problems from the perspective of virtue ethics. The closest and most obvious account is Petersen's approach (2007, 290–291). Petersen also briefly engages with virtue ethics and, however, he worries about the well-being and character of the designee, while Chomanski focuses on the well-being and character of the designer. Somewhat more distant and less obviously similar accounts are those that use virtue ethics in examining ethical problems with robots that are not persons. In particular, Rob Sparrow (2017, 2021) claims that virtue ethics enables us to clarify what is ethically objectionable in a situation in which a human being treats a robot which is not a person violently (e.g. by raping it), even if the robot is not—and cannot be—harmed. This objectionability results from the fact that this behaviour is an expression of vice: a negative trait in one's character. It is clear that it is similar to Chomański's approach, in the sense that virtue ethics enables us to condemn some actions taken towards robots, even if the robots do not suffer any harm because of them. Yet, the crucial difference is that Sparrow focuses solely on the character of the *users* of already existing robots, while Chomanski concentrates on the character of the *designers* of robots that are not yet built. In other words, Sparrow shows that it is bad for one's character to *use robots* in particular ways, while Chomanski demonstrates that it is bad for one's character to *design robots* in a specific manner. To be sure, Sparrow emphasizes the role of designers and the importance of the process of designing robots, yet he does that in terms of the duties of designers and the consequences of their actions rather than in terms of their character traits (2017, 474–475). Hence, Chomanski's position seems to be novel and specific in terms of focusing on the character of robot designers, and particularly on the vice of manipulativeness. In the next section, I focus only on the presence of manipulativeness on the side of designers in the cases of various types of APs.

4 Manipulativeness in designing various types of APs

The idea that the discussion about AP servants actually refers to a broader range of cases is not entirely new. For instance, Petersen, Musiał and Danaher have suggested that many issues connected to designing AP servants are in many relevant respects similar to the discussion of genetically enhancing human beings. Hence, many questions that appear regarding designing APs can be extrapolated to cases of genetic enhancement and vice versa. However, in this

section, my aims are more modest: I only examine which kinds of APs inevitably involve manipulativeness as a part of their design.

4.1 AP servants

Chomanski directly addresses the question of whether it is possible to design AP servants without manipulativeness, and I believe that he is right when he claims there is no way to design AP servants without the designers being manipulative (2019, 1011–1012). However, it is important to note that the kind of APs that Chomanski discusses is one of many, so it is at least possible that the APs discussed by Chomanski do have to be designed with manipulativeness, but some others do not. First, it is worth examining what exact features Chomanski ascribes to the AP servants that he discusses.

Chomanski initiates his investigations by asserting that AP servants should necessarily meet two conditions: the human-level AI condition (they should have "human-level intelligence, (…) autonomy, rationality, the possession of a moral character and a conception of the good" (Chomanski 2019, 995)) and the servitude condition (they "will have to be programmed in such a way as to be reliably willing to serve the relevant human needs (…) above almost all else" (Chomanski 2019, 996)). Hence, the first crucial feature of APs discussed by Chomanski is that they are artificial entities that possess all the necessary qualities to be considered a person, whatever these qualities are (Chomanski emphasizes that one of them should be human-level intelligence). This approach is common to all discussions about the APs mentioned above and is shared in my considerations presented below and the reason for this is obvious: if we want to discuss any group of artificial persons, they need to possess the features that persons possess (whatever these features are).

As for the servitude condition, it differentiates the kind of APs that Chomanski discusses from other kinds of APs. This condition makes the APs discussed by Chomanski servants and seems to refer both to—to refer to Petersen's distinction—"general servitude" and "specific servitude" (Petersen 2017, 289–291) (actually, Chomanski in his more recent text also distinguishes between these two kinds of servitude, although without using such terms (2021, 185)). General servants are those who do not have desires for any particular servitude actions, but for general servitude toward human beings, whatever they ask for—they simply have a "general" desire to serve humans. Specific servants are those who have a desire to do some particular tasks, such as a desire to wash clothes, iron them, etc. However, it is worth mentioning that specific servitude refers not only to APs designed to perform mundane tasks clearly being a kind of servitude, such as washing or ironing. It also refers to those that would be designed to be piano players or mathematicians. Moreover, it needs to be emphasized that both Petersen and Chomanski assume that the intensity of desires—either general or specific—to serve would be analogous to human beings' desire for food and water. Hence, both general and specific AP servants are designed to be "willing to serve the relevant human needs (…) above almost all else" and thus are meeting Chomanski's servitude condition.

It seems that in the case of both general and specific AP servants, the manipulativeness of designers is clear. Designers implement a general desire to serve or particular desires to perform some serving tasks and hence are "willing to orchestrate another's choices in an excessive manner" by limiting another's options and disregarding their rational capacities. Designers simply want both general and specific AP servants to do what they have been programmed to without taking other options into account or reasoning whether it is good to do so. Hence, both cases are clear examples of manipulativeness.

There are two main reasons why general and specific AP servants are designed with inevitable manipulativeness in mind. The first is the strength of desires that are assigned to them. The second is a lack of alternative—not designed by the programmers—desires they could obtain and follow. These two features limit APs' options and express a disregard for their rational capacities and therefore enable programmers to believe they can lead AP servants to act as they want them to. Hence, it is worth considering cases of APs in which desires programmed by designers are weaker and can be accompanied by desires not programmed by designers.

4.2 AP workers

One such case is described in Chomański's more recent paper, in which he discusses the case of APs made for profit (Chomanski 2021)—I will call them AP workers. Clearly, AP workers would meet the human-level AI condition, and they have all the features necessary to consider them persons equal to human beings. They would be designed to be able to participate in a labor market, but also—as Chomanski suggests—"a life outside of their work" (Chomanski 2021, 192). They could neither be owned by anyone nor be designed with a desire to cause suffering. Such AP workers could be hired by various companies that would pay the designers of AP workers for training them, analogous to how teams who have trained football players receive some money from their first employers. Finally, they would not be servants, either general or specific. What Chomanski means by this is that AP workers would possess neither a general desire to serve others unconditionally nor a narrow range of desires to perform certain particular tasks. Instead, they would possess a broad range of abilities, particularly the ability to learn. Moreover, AP workers would not be uncritical of their employers and their given tasks.

While AP workers would have to be designed to desire to work to some degree—otherwise, Chomanski's whole idea of APs made for profit would be pointless—this desire would not be as strong as in the case of AP servants. Chomanski mentions that AP workers could not only choose which company they would like to work for but also quit their job at a particular company to move to another, or they could also resign from working altogether (2021, 190). Moreover, Chomanski asserts that AP workers would also have "a life outside of their work" (Chomanski 2021, 192). While this statement is not elaborated by Chomanski, it seems reasonable to claim that this means that AP workers would have some desires that not only would not refer to work but also would not be programmed by the designers but would be acquired by AP workers themselves during their existence.

Before getting to the question of whether AP workers can be designed without manipulativeness on the side of designers, it is worth discussing the concept of AP workers in more detail, since it may raise understandable doubts and skepticism. Again, to imagine such a situation, we can recall the robot Andrew from "Bicentennial Man", who earns money by building and repairing clocks and is able to pay a human engineer to improve his embodiment with his own funds. Actually, a similar scenario is also debated outside the science-fiction realm. Samir Chopra with Laurence F. White (2011, 162–170) and Jacob Turner (2019, 197–201) discuss providing AI with legal personhood and, as a result, with a right to own property and money, to conclude contracts, or to sue others. Yet, one could wonder whether it is the right thing to do. Even assuming that we would agree that APs are persons in the moral sense, why would we decide that they are also persons in the legal sense and grant them powers such as those mentioned above? Why would people decide that robots can earn money and own goods? While there are various arguments for and against such a resolution, one possible answer is that if we recognize APs as moral persons equal to humans, and design them to have their desires and require some resources to maintain—such as electricity—and they were nobody's property (since it is rather uncontroversial that moral persons should not be owned by anyone), then letting them work and earn is probably the best way to enable them to achieve those desires and obtain the resources. Otherwise, we would have to provide them with electricity and other goods for free, which seems even more controversial, or to ignore their desires and needs, which would also be highly questionable, since we would be the ones who brought them into a miserable existence. Then, in a sense, granting robots moral personhood and making them equal to humans results in problems that can be solved by granting them legal personality. Such a scenario of course remains highly speculative and even more controversial. Yet, to repeat, one of the side- or meta-aims of this study is to signal the problems that would appear if robots became persons, and hence to enable us to decide whether developing such robots is an aim we would like to try to achieve at all.

Let us get back to the bottom line: Does designing such AP workers involve a manipulativeness on the side of designers? I believe it does. One could claim that a manipulativeness is not present because the desires programmed by designers for AP workers are significantly weaker than those programmed for AP servants, and also because AP workers can have their own desires not programmed by the designers. I assert that this only means that the manipulativeness is weaker, and not that it is utterly absent. Moreover, as Chomański shows in his "manipulativeness without manipulation" claim, when we discuss a designer's manipulativeness, we do not necessarily have to consider whether anyone is actually manipulated as a result or the consequences of manipulativeness in general. To remind, manipulativeness means "willing to orchestrate another's choices in an excessive manner" by "limiting another's options" and "disregarding their rational capacities." I contend that in the case of AP workers, designers attempt to lead APs to act as they want them to—they want to receive money from the company, and it is possible only if APs want to work and do so efficiently, so they design AP workers in such a way. In this sense, they aim to limit APs' options and disregard their rational capacities to shape their own future, though to a lesser degree than in the case of AP servants, but still. Actually, my claim that manipulativeness is present even if the designed desires are not overwhelming and accompanied by other, non-designed desires seems to be compatible with Chomanski's approach—he asserts that manipulativeness is present even if designers "do not exercise complete control over the AI's psychological profile" (2019, 1012). However, one could formulate at least three doubts about the claim that designing AP workers is manipulative.

Regarding the first doubt, one could claim that designing AP workers does not have to necessarily involve manipulativeness and that to eliminate it, it is enough to design weaker desires and leave more room for other desires. In other words, one could argue that the only reason why the case of AP workers involves manipulativeness is that the desires programmed by the designers are too strong and that they influence the APs too much in comparison to their other desires not programmed by the designers. As a result, to eliminate manipulativeness, it is sufficient to weaken the former and increase room for the latter. I disagree with this. I believe that in such a case, designers are still "willing to orchestrate another's choices in an excessive manner." Actually, the fact that they are willing to orchestrate another's choices is unequivocal. What might be equivocal is whether they do so in an excessive manner. I believe that they do so, since they attempt to significantly limit APs' options by disregarding their rational capacities. In particular, the

disregard for rational capacities is excessive, since there are ways to limit others' options without disregarding their rational capacities, as I show in the next paragraph. Moreover, I believe that additional excessiveness stems from the fact that such orchestration is not unavoidable, since it is possible to design APs without intentionally determining their future and orchestrating their choices—I discuss such cases in the next section. Therefore, I contend that if (1) we can make attempts to program some desires into APs without disregarding their rational capacities and (2) we can bring APs to life without intentionally determining their desires, then ignoring these opportunities results in an excessive willingness to orchestrate another's choices and, hence, involves manipulativeness.

Regarding the second doubt, one could argue that the understanding of manipulativeness demonstrated above is too broad, particularly because parenting and education turn out to be manipulative in light of it. After all, parenting and education are not unavoidable and may involve limiting options and disregarding rational capacities. This means that parenting and education can involve manipulativeness but do not have to. In particular, parenting and education can respect the rational capacities of the entity in question. Parenting and education can be about persuading and explaining, which respect the rational capacities of a child, but can also be about forcing and ordering, which disregard the rational capacities of a child. The latter case involves manipulativeness, while the former does not. Therefore, my understanding of manipulativeness covers some parenting and education cases in which a child's rational capacities are not respected. However, parenting and education that respect the rational capacities of a child are not manipulative, even though they may lead to a limitation of options and are not unavoidable. Simply speaking, parenting and education may be free of manipulativeness. In contrast, I believe that designers who choose to intentionally design APs do not—and actually cannot—respect the rational capacities of the APs they design. They do not raise or teach APs by asking, persuading, and explaining some things. They simply program particular desires without asking, persuading or explaining—without taking APs' rational capacities into account. This is a crucial difference between the intentional design of APs on the one hand and parenting and educating on the other. This is why I believe that while parenting and education—of both human beings and APs—do not have to involve manipulativeness, intentionally designing APs is unavoidably manipulative.

Regarding the third doubt, one could assert that in some cases, manipulativeness can prevent some significant harm or create some significant good and that in such cases, manipulativeness could be justified. While this might be true, this neither means that in such cases, manipulativeness is absent nor that it is not objectionable—it only means it can be justified. For instance, one of the conditions that Chomanski proposes AP workers should meet is designing them not to desire to intentionally cause another's suffering. Such designing involves being manipulative (since it expresses a "willing to orchestrate another's choices in an excessive manner" by "limiting another's options" and "disregarding their rational capacities"), yet this manipulativeness can be justified. Recall that in this study, I do not claim that the presence of manipulativeness either always or sometimes should directly result in condemning or banning an action that involves it. My aim is only to discuss in which cases of designing APs manipulativeness is present and in which cases it is not. Hence, to return to the bottom line, if the intentional design is the problem, then probably the best way to avoid the presence of manipulativeness is to develop APs without intentionally designing their desires. The following section discusses what this may look like.

5 Designing APs without manipulativeness: AP randoms

There seem to be two alternatives to the intentional design of AP desires. The first is to not design any "innate" desires at all, and the second is to randomize the selection of APs' desires, making it nonintentional. I will call APs resulting from these two modes of designing "AP randoms" to emphasize their main feature, which is their random character. Now, I briefly describe both modes of the non-manipulative design of AP randoms and I then discuss various dilemmas connected to bringing AP randoms to existence.

The first way of avoiding an intentional design of desires—and hence avoiding manipulativeness—is to not design any desires at all. AP randoms resulting from such a design would be "tabula rasa" with regard to desires. However, such an approach can be problematic if we recognize that some visions of personhood might consider "innate" desires a constitutive feature of personhood. In other words, from some perspectives on personhood, one cannot be a person without possessing some "innate" desires. If these were some specific desires, they could be intentionally programmed into APs and probably would not involve manipulativeness, since they are unavoidable in terms of being a necessary condition of regarding APs as persons. If these were some desires in general, they could be programmed by following the second approach to avoiding the intentional design of desires. This means that both methods can be combined, at least to a degree.

The second way of avoiding the intentional design of desires—and hence avoiding manipulativeness—is to randomize the selection of desires programmed into APs. Again, such randomization would refer to all the desires of APs apart from those that would be seen as constitutive of their

personhood. APs' desires would be randomly selected by an algorithm, which would work in a way that the designers could not predict (otherwise, the designers would still be manipulative). Moreover, the algorithm would not possess the status of a person (otherwise, the algorithm itself would be guilty of manipulativeness). Actually, Petersen mentions such a possibility when he examines what kinds of buttons can be pushed on a hypothetical machine able to create persons—Person-O-Matic—naming them "typical persons" (Petersen 2017, 286–287).

Apart from implementing one or both methods of non-intentional design, it is crucial to make such APs capable of changing their desires, at least to the degree that is given to human beings. It is important because otherwise, designers could still be accused of manipulativeness. They would excessively orchestrate APs' choices, limit their options and disregard their reasoning by making them unable to change their desires in any way. Hence, such APs with random desires would have to be able to understand, critically reconsider and transform their various features to free designers from being accused of being manipulative. Actually, Chomanski and Petersen consider a very similar ability a condition of the personhood of APs and refer to this as "a capability of reflection." Chomanski follows Petersen, who in turn follows Harry Frankfurt and contends that "It is plausibly a necessary condition of personhood that one be able to reflect on one's desires, for example, and reconsider them" (Petersen 2011, 289). Chomanski follows Petersen's account of reflection and asserts that reflection is a necessary condition for autonomy: "AI servants would be autonomous (…) provided that they are given the capacity to genuinely reflect on, and possibly change, their motivations, etc. (Chomanski 2019, 1003))." In turn, I believe that programming a capacity for reflection into AP randoms would additionally enable their designers to avoid manipulativeness.

It is worth noting that taking all of these measures against manipulativeness would have significant consequences. Specifically, developing AP randoms would result in creating APs who are unpredictable and uncontrollable at a level comparable to that of human beings. Such APs would no longer more or less faithfully achieve aims designed by humans but would follow their own—unpredictable and uncontrollable—aims, which might be incoherent with humans' aims. Most likely, the most radical example of this is the fact that in the randomized selection of AP randoms' desires, to avoid manipulativeness, we would have to enable the algorithm to choose features we might consider unethical. For instance, one could argue that enabling the algorithm to program into AP a desire to cause suffering to others would be wrong. However, eliminating such a desire from the range of available features might be considered manipulative, since it fulfils the features of being "willing to orchestrate another's choices in an excessive manner": it limits APs' options, disregards rational capacities and is not necessary. Of course, one could argue that manipulativeness is far less important than reducing suffering and that the latter can justify the former. Moreover, the fact that we would choose to avoid manipulativeness and enable APs to possess "innate" desires to cause suffering does not mean that we would simply accept this desire and its consequences—we could try to prevent it through socialization, as we do in the case of human beings who seem to share similar desires. Of course, the efficiency of such attempts would depend on the quality of socialization, the strength of desires and many other factors. Hence, the question is whether we want APs to be unable to hurt others or to be able to hurt others but choose not to do so. However, whatever we would choose to do with particularly controversial desires such as the desire to cause another's suffering, the bottom line here is that AP randoms developed in both aforementioned ways would be unpredictable and uncontrollable to a degree similar to human beings.

The obvious question is, what would be the point of developing AP randoms, and is it worth developing them at all? The general answer seems to be that the role of AP randoms in our lives would be analogous to the role of other human beings. As, somewhat romantically, Walker puts it, "we should create this sort of AI because we intend to love them, and hope that they love us; just as we intend and hope the same for our human children" (Walker 2006). Petersen also, while discussing "typical persons," argues that they "could plausibly bring ethical benefit to a great many couples who are not otherwise able to have biological children" (Petersen 2017, 287). John Danaher takes this idea one step further and presents the interesting and controversial idea of artificial offspring, who—as he sees it—could be in some ways better than natural offspring (Danaher 2017b).

Such ideas of embracing AP randoms involve a significant shift in our attitude toward technological artifacts. Of course, ideas of embracing any APs or providing robots with rights also involve them, but the case of AP randoms seems so radical that it illustrates the consequences of such embracing in a particularly stark and clear manner. In brief, this seems to illustrate that embracing AP randoms in particular and, to a lesser degree, APs in general involves some shifts in our perception of technological artifacts and our relations to them. First, and probably most generally, AP randoms would not be tools made to be used for humans' own good and to achieve humans' aims (as in the case of other technological artifacts), but persons created to be with humans who follow their own aims and try to achieve their own benefit. It is worth emphasizing that the good and aims of AP randoms may not be coherent with humans' good and aims, and sometimes the former may remain in contrast and compete with the latter. Moreover, as I have already mentioned,

AP randoms would be uncontrollable and unpredictable (to a greater degree than other types of APs and in stark contrast to other technological artifacts). This means that AP randoms would not be typical technological artifacts created to make humans' lives easier (however, while the intentional purpose of most technological artifacts is to make humans' lives easier, it is not rare for unintentional consequences of their implementation to have the opposite effect) but rather would make our lives more complicated (which, of course, does not mean "worse").

To be clear, the fact that we *can* design APs without manipulativeness by developing AP randoms does not mean that we *should* do so. First, while manipulativeness may always be objectionable, it might also be justifiable. Hence, it is plausible to say that its presence might be justified by defending some other values or preventing some other harms. In other words, while AP randoms seem to be the best (or even the only) means to design APs without manipulativeness, it does not matter that they are generally the best means to design APs. Second, and more generally, designing AP randoms in particular and, to a lesser degree, APs in general involves the abovementioned shifts in our attitude towards technology—from a controllable and predictable tool that serves our aims and makes our lives easier to an uncontrollable and unpredictable person who has its own aims, tries to achieve its own good, often competing with humans beings in doing so, and hence makes our lives more complicated. Whether we should make these shifts and elevate the status of technology is a matter of very detailed debate and by no means can be decided or even examined here.

Hence, to repeat, while there is a reasonable scenario in which we could design APs without being manipulative, it is by no means the only reasonable scenario on the table. There are also scenarios in which we can consider manipulativeness non-objectionable or objectionable but justifiable and develop not only AP randoms but also AP workers and AP servants. There is a possible scenario in which we could develop AP workers and AP servants instead of AP randoms (probably to avoid the unpredictability and uncontrollability of the latter). Finally, there is a scenario in which we should not develop any APs at all.

Advocating for any of these scenarios requires taking a position on the aforementioned shifts in our attitudes toward technology. Clearly, such a position does not have to involve a binary "yes or no" decision, but it may agree or disagree with the shifts to some degree. This of course involves examining many issues—not only manipulativeness—connected to APs' good (e.g., their autonomy) and humans' good (for instance, perspectives of technological unemployment or the meaningfulness of human life) and perspectives on their hierarchy ranging from anthropocentric humanism to anti-anthropocentric posthumanism. Again, all of this is a matter of a detailed debate, and examining this here is beyond the scope of this study.

6 Conclusion

In this study, I have discussed Chomanski's claim about manipulativeness to show that it significantly shifts the whole discussion on AP servants on the one hand and can be extrapolated to other kinds of APs on the other. Specifically, I have asserted that any design that intentionally determines AP's desires involves manipulativeness. Moreover, I have examined a way of designing APs that enables avoiding the presence of manipulativeness and have briefly discussed its consequences.

Funding Not applicable.

Code availability (software application or custom code): Not applicable. This manuscript has not been published and is not under consideration elsewhere.

Availability of data and material (data transparency) Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Baron M (2003) Manipulativeness. Proc Address Am Philos Assoc 77:36–54

Bloom P, Harris S (2018) It's Westworld. What's Wrong With Cruelty to Robots? The New York Times, https://www.nytimes.com/2018/04/23/opinion/westworld-conscious-robots-morality.html Accessed 20 November 2021

Bringsjord S, Naveen SG (2018) Artificial Intelligence. The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Zalta EN, Nodelman U (eds.), https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence. Accessed 2 September 2022.

Chomanski B (2019) What's wrong with designing people to serve? Ethic Theory Moral Prac 22:993–1015. https://doi.org/10.1007/s10677-019-10029-3

Chomanski B (2021) If robots are people, can they be made for profit? Commer Implicat Robot Personhood AI Ethics 1:183–193. https://doi.org/10.1007/s43681-020-00023-2
Chopra S, White LF (2011) A Legal Theory for Autonomous Artificial Agents. MI, University of Michigan Press, Ann Arbor
Columbus S (Director) (1999), Bicentential Man [Film]. Columbia Pictures
Danaher J (2017a) The symbolic-consequences argument in the sex-robot debate. In: Danaher J, McArthur N (eds) Sex robots: philosophical, ethical, and social implications. MIT Press, Cambridge MA, pp 79–99
Danaher J (2013) Is there a case for robot slaves? Philosophical Disquisitions (entry on blog on 22 April, 2013). https://philosophicaldisquisitions.blogspot.com/2013/04/is-there-case-for-robot-slaves.html. Accessed 20 October 2021
Danaher J (2017b) Why we should create artificial offspring: meaning and the collective afterlife. Sci Eng Ethics 24: 1097–1118. https://doi:https://doi.org/10.1007/s11948-017-9932-0
Danaher J (2019) The Ethics of Designing People: The Habermasian Critique. Philosophical Disquisitions (entry on blog on 19 April 2019). https://philosophicaldisquisitions.blogspot.com/2019/04/the-ethics-of-designing-people.html. Accessed 20 October 2021
Gellers J (2021) Rights for Robots: Artificial Intelligence, Animal and Environmental Law. Routledge, Abingdon and New York
Gordon JS (2020) What do we owe to intelligent robots? AI & Soc 35:209–223. https://doi.org/10.1007/s00146-018-0844-6
Gordon JS (2021) Artificial moral and legal personhood. AI & Soc 36:457–471. https://doi.org/10.1007/s00146-020-01063-2
Gordon JS (2022) Are superintelligent robots entitled to human rights? Ration 35(3):181–193. https://doi.org/10.1111/rati.12346
Gordon JS, Gunkel DJ (2021) Moral Status and Intelligent Robots. South J Philos 60(1):88–117. https://doi.org/10.1111/sjp.12450
Gunkel DJ (2018) Robot rights. MIT Press, Cambridge
Gunkel DJ, Wales JJ (2021) Debate: what is personhood in the age of AI? AI & Soc 36:473–486. https://doi.org/10.1007/s00146-020-01129-1
Mamak K (2022) Humans, Neanderthals, robots and rights. Ethics Inf Technol 24:33. https://doi.org/10.1007/s10676-022-09644-z
Musiał M (2017) Designing (artificial) people to serve–the other side of the coin. J Exp Theor Artif Intell 29(5):1087–1097. https://doi.org/10.1080/0952813X.2017.1309691
Neely EL (2014) Machines and the moral community. Philos Technol 27(1):97–111 (2014). https://doi.org/10.1007/s13347-013-0114-y
Nyholm, S. (2020) Humans and robots: Ethics, agency, and anthropomorphism. Rowman & Littlefield, London; New York
Petersen S (2007) The ethics of robot servitude. J Exp Theor Artif Intell 19:43–54. https://doi.org/10.1080/09528130601116139
Petersen S (2011) Designing people to serve. In: Lin P, Bekey G, Abney K (eds) Robot ethics. MIT Press, Cambridge MA, pp 283–298
Petersen S (2017) Is it good for them too? Ethical concern for the sexbots. In: Danaher J, McArthur N (eds) Sex robots: philosophical, ethical, and social implications. MIT Press, Cambridge MA, pp 155–171
Rini R (2017) Raising good robots. Aeon. https://aeon.co/essays/creating-robots-capable-of-moral-reasoning-is-like-parenting. Accessed 19 December 2021
Schwitzgebel E, Garza M (2015) A defense of the rights of artificial intelligences. Midwest Stud Philos 39(1):98–119. https://doi.org/10.1111/misp.12032
Sparrow R (2017) Robots, rape and representation. Int J Soc Robot 9(4):465–477. https://doi.org/10.1007/s12369-017-0413-z
Sparrow R (2020) Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained? Int J Soc Robot 13:23–29. https://doi.org/10.1007/s12369-020-00631-2
Tavani HT (2018) Can social robots qualify for moral consideration? Reframing the question about robot rights. Information 9(4):1–16. https://doi.org/10.3390/info9040073
Turner J (2019) Robot rules: regulating artificial intelligence. Palgrave Macmillan, Cham
Walker M (2006) A moral paradox in the creation of artificial intelligence: Mary Poppins 3000s of the world unite! In: Metzler T (ed) Human implications of human-robot interaction: papers from the AAAI workshop. AAAI Press, Menlo Park, pp 23–28
Walker M (2016) Free money for all: a basic income guarantee solution for the twenty-first century. Palgrave Macmillan, New York