AI and Ethics (2023) 3:1229–1239
https://doi.org/10.1007/s43681-022-00231-y

ORIGINAL RESEARCH

A new control problem? Humanoid robots, artificial intelligence, and the value of control

Sven Nyholm[1]

Received: 7 September 2022 / Accepted: 24 October 2022 / Published online: 7 November 2022

Abstract
The control problem related to robots and AI usually discussed is that we might lose control over advanced technologies. When authors like Nick Bostrom and Stuart Russell discuss this control problem, they write in a way that suggests that having as much control as possible is good while losing control is bad. In life in general, however, not all forms of control are unambiguously positive and unproblematic. Some forms—e.g. control over other persons—are ethically problematic. Other forms of control are positive, and perhaps even intrinsically good. For example, one form of control that many philosophers have argued is intrinsically good and a virtue is self-control. In this paper, I relate these questions about control and its value to different forms of robots and AI more generally. I argue that the more robots are made to resemble human beings, the more problematic it becomes—at least symbolically speaking—to want to exercise full control over these robots. After all, it is unethical for one human being to want to fully control another human being. Accordingly, it might be seen as problematic—viz. as representing something intrinsically bad—to want to create humanoid robots that we exercise complete control over. In contrast, if there are forms of AI such that control over them can be seen as a form of self-control, then this might be seen as a virtuous form of control. The "new control problem", as I call it, is the question of under what circumstances retaining and exercising complete control over robots and AI is unambiguously ethically good.

Keywords Control · Artificial intelligence · Humanoid robots · Self-control · Extended agency · The control problem

In late August of 2021, Elon Musk presented his plans for the "Tesla Bot". When he presented his ideas, Musk talked about the self-driving cars that Tesla has developed. He said that they are "basically semi-conscious robots on wheels", and that Tesla "is really the world's largest robotics company". Therefore, it makes sense, Musk continued, to give these robots a "human form". Hence the Tesla Bot.[1]

During his presentation, Musk explained some of the envisioned technical aspects of the Tesla Bot, as well as the motivation behind the robot: it could take over boring and repetitive tasks, so that humans don't need to do that work anymore. What is more interesting for the purposes of this paper, however, are some other remarks that Musk made about the Tesla Bot.

Those other remarks are especially interesting when one keeps in mind that Musk is worried that human beings might lose control over AI. He said, firstly, that the Tesla Bot would be "friendly". This seems to be a reference/allusion to Eliezer Yudkowski's [1] idea of "human-friendly AI", also known as "value alignment". Second, Musk said that the Tesla Bot would be weak enough that one could easily overpower it. Third, Musk said that the Tesla Bot would be slow enough that one could simply run away if one becomes scared. This fits with Nick Bostrom's [2] idea of "capability control": limiting the capacities of AI systems as a way of controlling them. As it happens, "value alignment" is another measure that Bostrom also discusses, which is also supposed to help us control AI systems. The idea is that if AI systems are human-friendly and aligned with our values, they are—at least to an extent—under our control.

In other words, Musk presented the Tesla Bot in a way that suggested that he had been thinking about how these robots can be kept under human control. What this example also brings out—as I see things—is that human control over

✉ Sven Nyholm
s.r.nyholm@uu.nl

[1] Utrecht University, Utrecht, The Netherlands

[1] A video of this presentation can be viewed here: "Elon Musk REVEALS Tesla Bot (full presentation)": https://youtu.be/HUP6Z5voiS8 (accessed on August 31, 2022).

AI is a problem with more ethical dimensions than those that people like Musk are worried about. The Tesla Bot has a human form, but it is designed in a way that gives it no individuality or no personality. It does not have a face; it does not have any clear personality. So, it is easy to instrumentalize this robot: to treat it like a tool—or like a slave, to paraphrase the title of Joanna Bryson's well-known [3] paper. Here we get to a potential problem: the idea of complete human control over AI becomes a more problematic notion if we create AI that takes the form of humanlike robots.

Imagine that the Tesla Bot did not look like it did, but that it looked and behaved even more like a human being, e.g., like Hiroshi Ishiguro's robot "Erica", which looks like a lifelike human woman. The idea of wanting to control—perhaps completely control—a robot that looks and acts like a human being can appear to be a problematic idea. So, even if we agree that control over AI is important, it is not clear that full human control over all AI systems is always a completely unproblematic goal or ideal.

My topic in what follows is the question of how we should think about the value of control and the different forms that AI can take. Is control always positively loaded? When it is good and positive, is the value of control always purely instrumental? Or is it the case that certain forms of control are also important or valuable in themselves, as ends and not only as means? Can control sometimes be bad or negative? Perhaps negative in itself? I will discuss the value of control—a topic that AI ethics researchers should discuss more often, and in more explicit terms—and I will relate it to different forms of AI agency. I will first reach some general conclusions about what forms of control are good, instrumentally and perhaps also in themselves, and some conclusions about what forms of control are bad/negative, either instrumentally or in themselves. Next, I will relate this to different ways in which we can think about the types of agents that AI systems can be seen as being, and the relation between their agency and our own human agency.

Briefly put, the positive side of my proposal is that if human control over AI can be conceptualized as a form of self-control, then control over AI is prima facie good, perhaps both instrumentally and intrinsically, i.e., both as a means and perhaps even as an end. On the negative side, my thesis is that if control over AI is, or can be seen as symbolizing, control over another person (or an entity that is a representation or a symbol of a human person), then control over AI can potentially be seen as negative, or at the very least as something that is in poor taste.

AI systems that can plausibly be viewed as pure "tools" are less interesting from this point of view. It will typically be instrumentally good to have control over them, and it will typically be risky or instrumentally bad to lose control over them. But like I said, I am here interested in the various different types of normative or evaluative status that we ascribe to different forms of control within human life. And I am interested in whether any forms of AI systems and their agency can be related to any non-instrumental ideas we might have about circumstances under which control can be seen either as in itself good or in itself bad. The *new control problem*, as I will call it, is the question of under what circumstances retaining and exercising complete control over robots and AI is unambiguously ethically good, and the challenge of separating those from circumstances under which there is something ethically problematic about wanting to have complete control over robots and AI.

1 Artificial intelligence and the standard control problem

Whenever there is talk about any form of AI or new technologies more generally, worries about control tend to come up. For example, one of the first things that many people—philosophers and others—tend to wonder and worry about in relation to self-driving cars is whether we will be able to retain enough control over them. The same applies to discussions about autonomous weapons systems. What if we lose control over them? [e.g., 4] In these kinds of cases, worries about control are often related to worries about potential responsibility gaps [5]. If the AI systems are operating autonomously, they will not be under our direct control, it is thought, and therefore it may be unclear who is responsible if there is an accident and somebody is hurt or even killed [6].[2]

Worries about control in relation to AI sometimes also arise from more general reflections on what AI is or should be taken to be [8]. Notably, when Alan Turing wrote some of his influential work on the topic of artificial intelligence in the early 1950s, the term "artificial intelligence" had not yet been introduced. Turing, instead, focused on the question of whether machines can think. That question, Turing suggested, is less clear than the question of whether machines can be made to behave in ways that imitate a thinking human being [9]. According to the so-called Turing test, we can say that we have invented an intelligent, thinking machine

[2] Such concerns are informed by real-life cases in which people have been harmed by these technologies [7]. For example, in 2018, it for the first time happened that a pedestrian was hit and killed by an experimental self-driving car. There was a safety driver in the car. But she seems to not have had enough control over the car to be able to prevent this accident from happening. In such cases, people often talk about "handing over control" of the operation of the car to the AI system in the car. We lose control, in other words, by giving it away. This seems problematic, partly because worries about potential responsibility gaps arise, but also partly because the idea of handing over control to an AI system is an uncomfortable thought for many people—especially if the AI is part of a technology that might be dangerous for human beings, such as cars or military robots.

if we invent a machine that can imitate a human being well enough that people cannot tell the difference between messages coming from the machine and messages written by a human.

On a Turing-inspired definition of AI, then, artificial intelligence is achieved when we create machines that can successfully *imitate* thinking human beings [10, 11]. The term "artificial intelligence" was coined just a few years later, in 1955, by a team of researchers in a research proposal for a famous workshop held at Dartmouth College in America. In that research proposal, there was not talk of "imitating" human thinking or intelligence. Rather, John McCarthy and his colleagues [12] spoke about *simulating* human thinking, learning, and other aspects of intelligence. Machines that simulate human intelligence are artificially intelligent, on this definition.

If we fast-forward to the 1990s and the first edition of Stuart Russell and Peter Norvig's [13] widely used textbook about AI—*Artificial Intelligence: A Modern Approach*—we can observe a shift in how AI is defined. According to Russell and Norvig, we should define AI as the creation of *intelligent agents*. An agent is here defined as a system that can "perceive" its environment and "act" in the pursuit of certain goals. Intelligence, in turn, is defined in terms of what philosophers call instrumental rationality: the capacity to efficiently achieve one's goals. Thus, the creation of artificial intelligence, on this way of thinking, consists in the creation of systems that can "perceive" their environment and "act" in that environment so as to efficiently achieve their goals. This could be a software agent—a computer program—or it can be a robot. AI, then, can be disembodied, so to speak, operating within a computer; or it can be embodied, operating in the natural environment.

Russell and Norvig define different types of artificial agents. For example, simple reactive agents can only act in response to specific stimuli that bring forth certain predetermined reactions. Agents operating with a model of the world, in contrast, can also use other information than their direct inputs—they can use a world model—to act in the service of their goals in more effective and dynamic ways. Yet even more advanced agents can also learn from their experience and become better able to achieve their aims over time, based on the training they get over time [13: 47–57]. For example, the computer program "AlphaGo" that was created by DeepMind was trained in two ways before it was able to beat Lee Sedol, who was the world champion of Go: it was trained on a huge set of data about actual Go games, and it was also trained by playing millions of games against itself [14]. In the end, the system was able to perform at a higher level than the human world champion. Agents can also be part of so-called multi-agent systems: different intelligent agents (which might include human agents) can join forces and work together in the service of goals the individual agents cannot achieve on their own [15].

When we get to these ideas of agents that can become better at achieving their goals over time, that can work together with other intelligent agents, and that can outperform humans at difficult tasks, we can see why worries about losing control over AI might arise. A common worry is that these systems will become more and more able to achieve whatever goals they have, and that we will not be able to retain control over these systems and their goal pursuit [8]. At some point, authors like Nick Bostrom [2] worry, we might even arrive at "super-intelligent" systems that are super-efficient at achieving whatever goals they are pursuing. This might lead to great risks—even so-called "existential risks" [16].

It is worth noting, however, that the tendency to have worries about a loss of control over AI is nothing new. Since the beginning of the field of AI, such worries have always been there. Turing, for example, already wrote the following in 1951 [17]:

> [I]t seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control. [18: 475]

The old AI control problem, then, existed even before the term "artificial intelligence" was invented! A few years later, in 1960, another pioneering AI researcher, Norbert Wiener, formulated the control problem in terms that are similar to those that are often used today. He wrote:

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively … we had better be quite sure that the purpose put into the machine is the purpose which we really desire. [19: 1358]

This formulation of the control problem brings up the idea of "value alignment", which was already mentioned above. When Stuart Russell [8] discusses this idea, he talks about what he calls the "King Midas Problem". King Midas had the wish that everything he touched would turn to gold. That seemed to him like a great idea. However, as soon as his food, his drinks, and his family turned to gold when he touched them, and he started starving because he could no longer eat, it soon became clear to King Midas that getting what he wanted was not such a good thing. The parallel to AI here is supposed to be that if we specify the goals that AI systems are intended to achieve in the wrong ways—or if the AI systems somehow misinterpret the goals they are supposed to achieve—then this can lead to disastrous consequences.

Here, lastly, is another, succinct formulation of the control problem, as many researchers working on this issue currently conceive of the problem. The philosophically minded computer scientist Roman Yampolskiy [20: 1] writes:

> [The] invention of artificial general intelligence is predicted to cause a shift in the trajectory of human civilization. To reap the benefits and avoid pitfalls of such powerful technology it is important to be able to control it.

In summary, handing over control—or losing control to—a system that is learning and highly efficient at achieving whatever goals it is pursuing strikes many researchers and others—from computer scientists, to philosophers, to regular people—as being risky. It can lead to bad consequences, perhaps even on a massive scale. But do these concerns about avoiding possible bad effects capture everything about the way in which we—or most of us—value control within our lives? And is control always an unambiguously good thing? And what do we mean by control to begin with?

I will now first make some observations about control, and then after that discuss some different ways of understanding the value/importance of control. To anticipate what is to come: I think that control is not something that is only ever valued as a means to the end of being safe. I think that we often also value control as either an end in itself or as a core element of things that have value as ends in themselves. I also think that control can sometimes have a negative value—including a non-instrumental negative value—and all of this is relevant, I suggest, to how we should think about the AI control problem. It also gives us reason to reflect more on how we think about the agency that we attribute to AI systems and how it relates to our own human agency.

2 What is control?

Control is discussed in different areas of philosophy, including but not limited to the philosophy of technology, general moral philosophy, and political philosophy [e.g., 4, 21–24]. Control is also, of course, analyzed and discussed within other fields, such as computer science and related fields that study topics like AI. There is also what is called control theory, an important and sophisticated part of the study of engineering [25]. My focus here is not on the design of control systems as conceived of within control theory, but rather on the idea of human beings controlling or exercising control over different things or phenomena. And I will primarily approach the issue of control from a philosophical point of view. I will not discuss the work that has been done on control in these different areas of philosophy and elsewhere in detail here, but instead extract what I think are some key lessons that one can learn when one looks at what people say about control within these different discussions. In particular, I wish to put forward three main observations about control, which I think find support in the overall literature on control.

Firstly, control can be more or less direct [22, 24]. For example, if one is driving a conventional car, one can control what direction the car is going by turning the steering wheel. If you want to go right, you can turn the wheel in that direction, and if you want to go left, you can turn the steering wheel in that direction, whereas if you want to go straight without turning, you can center the steering wheel. This is a fairly direct form of control over what direction the car is traveling in. In contrast, if you are being driven around by a driver who is willing and able to follow your instructions, you can in a more indirect way control in what direction the car is going by asking your chauffeur to drive in the direction you would like to go. Or, to take another example, under normal circumstances, you can directly control whether or not you are raising your arm by deciding whether or not to raise it, and then simply going ahead and raising it or not. In contrast, you have no similar form of direct control over your mood. If you are in a bad mood, you cannot directly will yourself into a good mood. However, there can be indirect ways of controlling your mood. If you know that jazz, a run in the forest, or a cup of hot chocolate tends to put you in a good mood, you can listen to some jazz, go for a run in a forest, or have a hot chocolate, and thereby indirectly control (or try to control) your mood [21].

Second, control can be more or less robust [27].[3] That is to say, depending on what it is that you are trying to control, you might be able to remain in control over it in a wider or perhaps a narrower range of different circumstances. If you are a skilled driver, for example, you can remain in control over your car both in good weather and in bad weather, both when you are tired and when you are alert, and so on. Somebody who is a less skilled driver might have less robust control over their car: e.g., they may only be able to remain in control over it in favorable weather conditions or when they are alert and not tired.

Third, control is a multi-dimensional thing [5]. There are many different aspects to control, and we can have more or less control along those different dimensions. Without going into too much detail, here are some of the different aspects or dimensions of control that one can find in the literature about control. Whether one has control over something involves, but may not be limited to, the following different aspects:

1: whether something aligns with, or tracks, one's values, wishes, or instructions
2: whether one understands a thing, and if so, to what extent and in what detail

[3] As one of the anonymous peer reviewers reminded me, the idea of robustness is also a key part of control theory within engineering studies, with a whole field devoted to "Robust and Optimal Control", as described, for example, in Tsai and Gu's [26] book with that name.

3: whether one is able to monitor what one is controlling
4: are there interventions that one can take, and if so, how precisely is one able to steer something, or how often and easily can one intervene?
5: is one able to change, update, or discontinue/stop something one is controlling?

When all of these—and any other—aspects or dimensions of control are all in the same hands, so to speak, and the person has a full measure of all of them, that person can be said to have a maximum amount of control over the thing in question, especially if their control is also very robust across a maximal range of different circumstances.

More typically, though, these different aspects of control may not all be maximally realized. They may also be spread across different people [28]. Moreover, the degree to which an individual or a group has access to these different aspects of control might be limited, and it might also not be very robust [5]. Control, then, admits of degrees, and it admits of degrees along a number of different dimensions.

When one thinks about control in the way outlined above, one can immediately see how it might be difficult to maintain complete control over certain forms of AI. Some AI systems will be "black boxes" to us, for example, because we cannot fully understand the patterns in the artificial neural networks in the AI systems [10]. The control we are able to have over the AI systems may also not be very robust. We might be able to control them in laboratory settings, i.e., in very controlled environments. But once the AI systems are operating in the "real world", it might be much harder to retain control over them along all the different dimensions of control [8]. And many different people might have some share, but perhaps only a limited share, in the different aspects of control in relation to some AI system. So, many different people might have some small measure of control. But no one might have maximal control. And the people might not be part of a well-run organization with a clear division of responsibilities.

In any case, the ideas above are some key aspects of the complex issue of what control consists in, as I understand it here. Much more can be said about the nature of control. But for now, I will leave it at that, and instead turn to the question of how we should think of the value/importance of control. In addressing that issue, I will also draw on discussions about control that one finds in different areas within philosophy.

3 The normative and evaluative status of control: What different kinds of value can control have?

As noted above, when people discuss control in the context of AI, the assumption often seems to be that having control will help to produce good effects, whereas losing control over AI can produce dangerous effects and risky situations. The value of control, in those discussions, is portrayed in a primarily instrumental way. Control is seen as a means to other ends, typically the ends of safety and security. There is even a growing interdisciplinary field called "AI safety and security", whose main focus is on how to achieve control over AI [20]. But if we zoom out a little and think about control more generally and how it matters in human life, it is less clear that control is only ever valued as a means to other ends. It then also becomes less clear that control is always something that has a positive value.

Regarding the idea that control might not always have a positive value, I will first quickly highlight an idea that Yampolskiy [20] presents in his recent work: namely, that having direct control over AI can sometimes be instrumentally bad from the point of view of safety (cf. [22]). To use an example Yampolskiy himself uses, if a self-driving car follows all of our orders in a direct way—e.g., by voice control—this might be dangerous. Suppose that somebody in a self-driving car on a highway tells the car to stop and the self-driving car directly follows all orders that humans issue to it. The car might then abruptly stop in the middle of the highway and cause a major crash. Yampolskiy discusses several similar kinds of cases, and argues that it might be impossible—at least in some circumstances—to have full control and complete safety at one and the same time.

That was a quick motivation for thinking that (complete) control is not always unproblematic. In a more thorough analysis of the value of control, the following question is useful to always keep in mind: *who is controlling whom or what*? Depending on who is controlling whom or what, control may have a positive value, and might even be seen as a good thing in itself, in a non-instrumental way. Alternatively, it might have a negative value, perhaps even a non-instrumental form of negative value.

Let's start with the positive side. Notably, different forms of *self-control* are sometimes seen as not only being good as a means to other ends, but as being good or valuable in themselves. Consider, for example, the view held by those who view self-control as a key aspect of virtue, and who also value virtue as a goal or an end in itself. On such a view, self-control has—at least in part—a non-instrumental positive value [29].

That view can be associated with virtue ethical views, such as the views defended by some ancient philosophers, like the Stoics. Others also make claims that point in this direction. At the beginning of the *Groundwork for the Metaphysics of Morals*, for example, Kant [30] writes that "self-control [Selbstbeherrschung] and careful deliberations can be seen as being part of the *inner worth* of a person". He goes on to add that this would be of conditional value, and only be unconditionally good if coupled with a good will. However, it is noteworthy that Kant views self-control as part of a person's "inner worth".

Along Kant-inspired lines, it is also possible to note that control over oneself can be seen as an aspect of personal autonomy, which is something that is also typically valued as being important in itself, and not only as a means to other ends [30]. Also of a broadly Kantian flavor is Jeremy Waldron's [31] claim that self-control is an aspect of human dignity. Human dignity, on Waldron's view and most other views, is important in itself. So, if self-control is part of dignity, that is another argument in favor of viewing self-control or control over oneself as something with non-instrumental value or importance.

We can also here consider Martha Nussbaum's [32] claim that "control over one's environment" is one of ten crucial capabilities that are part of a good and dignified life, where these capabilities are claimed to be good as ends, and not only as means. Control over one's environment is spelt out by Nussbaum as involving the right to political participation and the right to work and to have property. Accordingly, Nussbaum's view of what the capability of control over one's environment amounts to is similar to some views about what it is to be an autonomous person. At any rate, control over one's environment is treated as good in itself, and not only as a means to other ends, on Nussbaum's capability theory.

Control over one's own body is sometimes also admired as something positive in itself, and not only as a means to other ends. Think about gymnastics, for example. When Simone Biles and other athletes compete in gymnastics, they receive high scores and are admired when they display great control over their own bodies. In general, different forms of virtuosity are often admired as valuable in themselves. And this also involves different forms of control, e.g., over an instrument (such as a violin). There is something impressive about—and it is typically viewed as a great achievement in itself to have attained—a high degree of control over one's body or an instrument. This is a form of mastery or virtuosity that we tend to admire in a partly non-instrumental way.

In contrast, if what somebody is trying to control—or is succeeding in controlling—is not themselves or an instrument, but *another person*, then the value of this control is radically different. One person controlling another person—e.g., having that other person as their slave, at the extreme—is usually seen as very negative and bad in itself ([23, 33], see also [34] on the badness of some employers' excessive control over their employees.) Being unfree, because one is under other people's control, is typically seen as being bad in itself. And somebody who tries to control other people is typically seen as acting immorally, since it is bad and wrong in itself to try to control other persons. It is one thing to try to exercise control over one's child or a non-human animal. This can even be seen as good. It is quite another to try to exercise control over some person who is a fully mature moral agent and who should be treated as one's moral equal. This is seen as bad.

The last observation I will make about the value of control is that control—while it can be good in certain ways, and sometimes even intrinsically good—is something towards which one ought to exercise a certain amount of moderation [22]. Some people are said to be "control freaks". The idea is then not, presumably, that they are making a mistake in wanting to have a certain amount of control. The mistake they are seen as making is, rather, that they want to have too much control over something. Control—e.g., self-control—can be a good thing and even good in itself, but it also seems possible to be too obsessed with control to a point where one can appropriately be labeled a "control freak."

Let us now assume that these ideas about control and its value are acceptable, at least roughly speaking. In particular, let us focus on the ideas that self-control is positive (perhaps even in a non-instrumental way) and that control over other people is negative (typically in a non-instrumental way). With these ideas taken on-board, let us relate them back to the issue of control over different forms of AI. If we accept the just-considered claims about the value of control, what does that tell us about the value of human control over AI? In now considering that question, I will be exploring different ways of thinking about the relation between people and the agency of the AI technologies they are using or interacting with.

4 The value of control and types of AI agency, part I: extended self-control

In general, the above-considered claims about the different ways in which control can be positively or negatively valuable motivate two general theses. Firstly:

> (1): if human control over certain forms of AI can be conceptualized as some form of self-control, then this control over the AI in question might not only be instrumentally good, but could potentially also in certain respects be non-instrumentally good.

Such control over AI could then be seen as some form of virtue, as a dignified way of controlling oneself, as part of one's personal autonomy, and perhaps as a key capability. The question arises, then, of whether there is any sensible way of understanding any instances of human control over AI as ever being a form of self-control. Before we get to that question, however, here is a second implication of what was discussed in the previous section:

> (2): if there are any AI agents that could be seen as persons in any important sense, or if any AI agents

> could represent or symbolize persons, then control over those AI agents might be an in itself bad or morally problematic form of control, just like one human person exercising control over another human person can be an in itself problematic thing.

In relation to any AI agents which might be regarded as being or representing some form of persons, we could say that this does not only create a new control problem, but also a control dilemma. Losing control over these AI agents that appear to be some form of persons might be problematic or bad because it might be unsafe, on the one hand. Having control over these AI agents might be morally problematic because it would be, or represent, control over another person, on the other hand. The crucial question here, though, is whether any AI agents could ever be persons or properly be seen as persons, or whether they could ever be representations of persons in any significant sense.

A lot depends on how we think about the agency of AI technologies. We saw above that when computer scientists—such as Russell and Norvig [13]—talk about AI systems, they take for granted that we can view AI systems as a form of agents. This has even become one standard way for computer scientists to define what an AI system is. To a philosopher, in contrast, whether an AI system is an agent might appear to be an open question [35: chapter two; 36]. This might be because the average philosopher has a more maximal or demanding view of what it is to be an agent, whereas the average computer scientist has a more minimal view of what it is to be an agent [37]. It is worth noting, for example, that even computer scientists who are very skeptical about anthropomorphism or the need to worry about things like superintelligence—e.g., the philosophically inclined computer scientist Virginia Dignum [10]—have no trouble defining or understanding AI systems as a form of agents. That helps to illustrate that by "agent", computer scientists mean something much less loaded than what the average philosopher means by it.

That helps to pave the way for the possibility that the agency of an AI system can be understood in very different ways. In particular, one possibility is to view the agency of AI systems as not being independent of human agency. Instead, at least with respect to some forms of AI agency, we could view AI agency as a form of extension of our own human agency [35, 38]. When we use AI technology, we could potentially be seen as acting through the AI systems we are creating. AI systems, on such a view, would be different than traditional non-agential tools (e.g., a hammer or a frying pan), because the AI systems would pursue goals and respond to their environments in ways that traditional tools cannot do. Yet, when we think about the goals and goal pursuit of the AI system, we might view these goals—as the philosopher Elena Popa [39] argues and as Stuart Russell [8] also argues that we should—as being our human goals. The goal pursuit of those AI systems can then be conceptualized as an extended form of human goal pursuit. Along such lines, we might think of our human goals as being extended out into some of our AI systems, and we might think of ourselves as acting through, or via, the AI systems we use. This would be one way of moving in the direction of thinking of human control over AI systems as a form of self-control. Specifically, one could take such a view if one understands human self-control as consisting, among other things, in control over one's agency—where this agency might include technological extensions of our human agency.

Relatedly, the philosophers José Hernández-Orallo and Karina Vold have recently argued that we can view the "thinking" that some AI technologies do as extensions of our own human thinking, so that these AI technologies become parts of our "extended minds", to use the expression made famous by Andy Clark and David Chalmers [40, 41]. If we view the information processing or reasoning done by certain AI systems as extensions of our human thinking in such a way, this could also pave the way for the possibility of conceptualizing human control over at least some AI technologies as a form—perhaps an extended form—of self-control. The idea would then be that control over one's own thinking/thinking processes is a form of self-control.

A question here is whether we should think of the human side of things as a form of *individual agency*, or whether the idea of *group agency* will perhaps more often make sense when we think of the use of AI as a form of extended agency [5, 42]. Human beings often use AI systems within the context of work they do within groups and organizations. The police in some district, for example, might use an AI system as part of their activities, or the staff at a hospital might also use an AI system in their work. Similarly, a tech company—such as a social media company—might use AI systems as part of how they run their websites or other services. In such cases, if these organizations keep their AI systems under their control, then this might be viewed as a form of group level self-control, or organizational self-control.

An organization's ability to maintain control over the way in which an AI system helps to pursue the organization's goals can be seen as a form of virtue on the part of that organization. That could then be seen as something that may not only be instrumentally good, but potentially also positive or admirable in itself. On the flipside, when an organization does not have full control over the AI systems that they use, they can be criticized for failing to exercise an appropriate level of organizational self-control. For example, it could be seen as a virtue—and a commendable form of extended self-control—if a military unit exercises control over military AI technologies they use, whereas it could be seen as a vice—a problematic lack of group level self-control—if they fail to exercise control over dangerous AI technologies that they use.

It could of course also be that an individual is seen as extending their individual agency by making use of AI technologies to achieve his or her goals. If that individual then loses control over that AI system in the pursuit of his or her goals, that might be seen as undignified and as a loss of personal autonomy on the part of the person. It might be a little like first drinking too much, getting into a car, and then losing control over one's car. That can be seen as a form of extended loss of self-control. In the same way, somebody who loses control over an AI system they are using to try to achieve their aims might also be seen as losing control over their own agency or part of their own agency.

There seem to be some potential ways, then, in which we could understand the use of AI systems, at least in some cases, as a form of extension of one's own agency or as an extension, on the part of a group or organization, of the group's organizational agency. That would pave the way for thinking of some instances of human control over AI in terms of the notion of self-control, so that the loss of control over the AI in those cases amounts to a loss or failure of self-control.

Such instances of human control over AI, which can be conceptualized as (extended) forms of self-control, can be seen as good and virtuous in themselves, since self-control is widely seen as non-instrumentally valuable in various ways, as noted in the previous section. But what about the other possible way of thinking about human control over AI agents highlighted above? That is, could it ever make sense to view AI agents as a form of persons or a representation of persons, who it might then be ethically problematic to try to control, since it is ethically problematic to try to control persons? [3] Let us now consider that question.

5 The value of control and types of AI agency, part II: humanoid robots

Notably, there are philosophers who argue that AI systems might one day become persons, whose moral status we should respect. Some philosophers even argue that we should potentially treat—or that it would not be a mistake to treat—some existing AI systems as a form of moral persons.[4] These authors typically focus on AI systems in the form of (humanoid) robots, rather than software agents or computer programs. Philosophers like Mark Coeckelbergh [43], David Gunkel [44], John Danaher [45], Janina Loh [46], Eric Schwitzgebel and Mara Garza [47, 48], and Chris Wareham [49] argue that some robots are or might become moral persons, to whom we owe some degree of moral consideration. While philosophers like Schwitzgebel and Garza argue that AI-equipped robots might become genuine moral persons because of their capabilities, others—like Coeckelbergh, Gunkel, Wareham, and Loh—argue that robots can become moral persons because of how we interact with them. Danaher, in turn, has a Turing-inspired view according to which robots should be treated like moral persons if they are able to imitate, or consistently behave like, moral persons.

My own view is that it makes sense here to ask not only about whether robots can have morally relevant abilities or imitate abilities that might make them into moral persons, but that it also makes sense—and perhaps more sense—to ask whether robots can be seen as *representing* or *symbolizing* moral persons ([35, chap. 8]; see also [50] on symbolic value more generally). Robots that are designed to look and behave like human beings, in particular, can be seen as a form of representation of, or symbol for, human beings [51, 52]. This will not make the robots themselves into moral persons to whom we owe moral consideration. But it can be enough to make it the case that it becomes ethically problematic or unfitting to treat the robots in certain ways, since certain forms of treatment of such robots might be seen as representing or symbolizing problematic ways of treating human beings [52]. If we perform acts of violence against robots made to look and behave like human beings, for example, this can be viewed as ethically problematic because it glorifies or glamourizes violence against human beings.[5]

If a robot can be a moral person, as some philosophers think, or if it can imitate or symbolize/represent a moral person, then any of those possibilities might make it problematic to want to have the robot under our full control. Personally, I am skeptical about the idea that any robots might have, or come to have, properties or abilities that would genuinely make them into full moral persons. In particular, I am skeptical about the idea—which Schwitzgebel and Garza [47, 48] take very seriously—that robots might come to have humanlike minds, with humanlike consciousness and emotions. But I grant them that if robots could come to have such minds, then they would potentially, for this reason, become full moral persons to whom we owe the same form of moral consideration that we owe to our fellow human beings. But that is a big "if" and not, in my view, a very realistic one. (Cf. [35, chap. 6]) What is more interesting, as I see things, is to focus on the possibilities that robots might imitate or represent/symbolize human persons with full moral status. Those possibilities are

[4] Relatedly, in June of 2022 the controversial Google engineer Blake Lemoine made headlines when he claimed that the large language model "LaMDA" had become a sentient person who should be given rights and moral standing.

[5] Similarly, Robert Sparrow [52] discusses this idea in the context of sex robots, and he argues that performing sex acts on a seemingly non-consenting sex robot represents or symbolizes non-consensual sex or rape in a highly morally problematic way, even if the robot itself lacks a mind and lacks moral status.

enough, in my view, for it to become morally problematic—or at least not completely unproblematic—to want to have such robots operating under our complete control. The reason for this would not be that it would be immoral towards these robots themselves, but rather that it would symbolize or represent something that is deeply morally problematic: namely, persons trying to control other persons.

At the same time, it would also be problematic—since it might be unsafe—to not have control over such humanoid robots. So, the best solution seems to be to avoid creating humanoid robots unless there is some very strong reason to do so that could help to outweigh the symbolic problems with having a humanoid robot that we are exercising complete control over. Or, alternatively, we might try to exercise control over these robots in a way that signals that we find it distasteful to do so, or that at least signals and acknowledges that we find it wrong to try to control real human beings.

The Tesla Bot, for instance, seems like an example of a case where it is unnecessary to give the robot a human form. Moreover, if future versions of the Tesla Bot are given a much more humanoid form and are made to display more humanlike behavior than what Musk originally described, then the kind of complete control over it that Musk envisioned in August 2021 might become morally controversial, since this could then be seen as a symbolically problematic representation or expression of a wish to have complete control over other persons, viz. something we think of as ethically problematic. In some other cases—e.g., the therapy robot “Kaspar”, which is used in experimental treatment of children with autism—it might make more sense to actually give the robot a humanoid form [53]. For therapeutic purposes, some robots might need to have a humanoid form. When we interact with those robots—e.g., Kaspar—we might avoid treating them in a way that appears to symbolize wanting to be in control over another human being. In contrast, if robots and other AI systems do not at all look like humans, and their behavior is not very humanlike at all, these kinds of issues and worries about it being somehow improper to want to control these robots do not arise in the same way.

6 Concluding discussion

Above, I have argued that we should not only think of control over AI systems in terms of instrumental value. We should also consider whether there might be forms of human control over AI that can be seen as non-instrumentally good as well as whether there might be forms of human control over AI technologies that could potentially be seen as ethically problematic, perhaps in a non-instrumental way. I have focused on two main forms of control, which differ radically in how they are typically evaluated in moral philosophy and beyond: self-control, on the one hand, and control over other persons, on the other.

Self-control is often valued as good in itself or as an aspect of things that are good in themselves, such as virtue, personal autonomy, and human dignity. In contrast, control over other persons is often seen as wrong and bad in itself. This means, I have argued, that if control over AI can sometimes be seen or conceptualized as a form of self-control, then control over AI can sometimes be not only instrumentally good, but in certain respects also good as an end in itself. It can be a form of extended self-control, and therefore a form of virtue, personal autonomy, or even human dignity.

In contrast, if there will ever be any AI systems that could properly be regarded as moral persons, then it would be ethically problematic to wish to be in full control over them, since it is ethically problematic to want to be in complete control over a moral person. But even before that, it might still be morally problematic to want to be in complete control over certain AI systems; it might be problematic if they are designed to look and behave like human beings. There can be, I have suggested, something symbolically problematic about wanting to be in complete control over an entity that symbolizes or represents something—viz. a human being—that it would be morally wrong and in itself bad to try to completely control.

For these reasons, I suggest that it will usually be a better idea to try to develop AI systems that can sensibly be interpreted as extensions of our own agency while avoiding developing robots that can be, imitate, or represent moral persons. One might ask, though, whether the two possibilities can ever come together, so to speak.

Think, for example, of the robotic copy that the Japanese robotics researcher Hiroshi Ishiguro has created of himself [35]. It is an interesting question whether the agency of this robot could be seen as an extension of Ishiguro’s agency. The robot certainly represents or symbolizes Ishiguro. So, if he has control over this robot, then perhaps this can be seen as a form of extended agency and extended self-control. While it might seem symbolically problematic if Ishiguro wants to have complete control over the robot Erica that he has created, which looks like a human woman, it might not be problematic in the same way if he wants to have complete control over the robotic replica that he has created of himself. At least it would be different in terms of what it can be taken to symbolize or represent.

Erica the robot, as far as I know, is not supposed to be a replica of any particular other real human being. But the robot does look extremely similar to a real human being. That might be seen as being enough for it to be morally problematic to want to be in complete control over that robot. However, it would be worse if somebody created a robotic replica of somebody else—i.e., of a particular person—and then wanted to be in complete control over that robot. The reader can imagine, for example, that a robotic copy of you is created and that the creator of that robot would then want to have complete control over that robot copy of you or that the creator would sell that robot replica of you to somebody else who would then exercise complete control over the robot. I suspect that many of us would feel very uncomfortable about that prospect. We don't like the idea of somebody making a robotic copy of us that they then want to have complete control over, because we do not like the idea of others wanting to have control over us or over something that looks and acts like us.

This suggests to me that if we create humanoid robots that we want to retain complete control over, those robots should not be made to look like any particular real people (other than perhaps ourselves!). And again, it might be best to avoid creating humanoid robots to begin with, since we can then avoid these kinds of worries about whether there is something symbolically or otherwise ethically problematic about wanting to be in complete control over these AI agents [cf. 3].

In conclusion, from the point of view of control and its value, the best AI systems that we can create seems to be ones that can be seen as extensions over our own agency, over which we can have control that can be viewed as a form of extended self-control. Such systems are not likely to take on a humanoid form. They are more likely to be computer programs (software agents) or robots with a non-human form.

Acknowledgements For helpful feedback, many thanks to John Danaher, Daniel Tigard, audiences at Utrecht University, LMU Munich, the ETEG Workshop, the Philosophy of Digitality Open Space, and the Robophilosophy 2022 conference, as well as the two anonymous peer reviewers.

Author's contribution SN is the sole author.

Funding This work is part of the research program Ethics of Socially Disruptive Technologies, which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO Grant Number 024.004.031).

Data availability This research was purely theoretical in nature and made use of no data that had been gathered or that could be shared.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Yudkowsky, E.: Artificial intelligence as a positive and negative factor in global risk. In: Bostrom, N., Ćirković, M.M. (eds.) Global Catastrophic Risks, pp. 308–345. Oxford University Press, New York (2008)
2. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Oxford (2014)
3. Bryson, J.J.: Robots should be slaves. In: Wilks, Y. (ed.) Close Engagements with Artificial Companions, pp. 63–74. John Benjamins, London (2010)
4. Santoni de Sio, F., van den Hoven, J.: Meaningful human control over autonomous systems: a philosophical account. Front. Robot. AI **5**, 15 (2018). https://doi.org/10.3389/frobt.2018.00015
5. Nyholm, S.: Attributing agency to automated systems: reflections on human–robot collaboration and responsibility-loci. Sci. Eng. Ethics **24**(4), 1201–1219 (2018)
6. Hevelke, A., Nida-Rümelin, J.: Responsibility for crashes of autonomous vehicles: an ethical analysis. Sci. Eng. Ethics **21**(3), 619–630 (2015)
7. Nyholm, S.: The ethics of crashes with self-driving cars: a roadmap, I. Philos. Compass **13**(7), e12507 (2018)
8. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Penguin, London (2019)
9. Turing, A.: Computing machinery and intelligence. Mind **LIX**, 433–460 (1950)
10. Dignum, V.: Responsible Artificial Intelligence. Springer, Berlin (2019)
11. Gordon, J.-S., Nyholm, S.: Ethics of artificial intelligence. Internet Encyclopedia of Philosophy. https://iep.utm.edu/ethic-ai/ (2021)
12. McCarthy, J., et al.: A proposal for the Dartmouth summer research project on artificial intelligence. Available here http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf. Accessed 17 Nov 2021 (1955)
13. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, New York (1995/2020)
14. Chivers, T.: How deepmind is reinventing the robot. IEEE Spectrum. https://spectrum.ieee.org/how-deepmind-is-reinventing-the-robot. Accessed 25 Oct 2021 (2021)
15. Weiss, G.: Multiagent Systems, 2nd edn. MIT Press, Cambridge (2016)
16. Ord, T.: The Precipice: Existential Risk and the Future of Humanity. Hachette Books, New York (2020)
17. Turing, A.: Can digital computers think? TS with AMS annotations of a talk broadcast on BBC Third Programme 15 May 1951, The Turing Digital Archive: http://www.turingarchive.org/browse.php/B/5. Accessed 31 Oct 2021 (1951)
18. Turing, A.: Machine intelligence: a heretical theory. In: Copeland, B.J. (ed.) The Essential Turing. Oxford University Press, Oxford (2004)
19. Wiener, N.: Some moral and technical consequences of automation. Science **131**(3410), 1355–1358 (1960)
20. Yampolskiy, R.: On controllability of AI. arXiv:2008.04071 (2020)
21. Debus, D.: Shaping our mental lives: on the possibility of mental self-regulation. Proc. Aristot. Soc. **116**(3), 341–365 (2016)
22. Di Nucci, E.: The Control Paradox. Rowman & Littlefield International, London (2020)
23. Pettit, P.: On the People's Terms. Cambridge University Press, Cambridge (2012)

24. Schmidt, A.: Domination without inequality? Mutual domination, republicanism, and gun control. Philos. Public Aff. **46**(2), 175–206 (2018)
25. Levine, W.S. (ed.): The Control Handbook. CRC Press, Boca Raton (2011)
26. Tsai, M.-C., Gu, D.-W.: Robust and Optimal Control. Springer, Berlin (2014)
27. Himmelreich, J.: Responsibility for killer robots. Ethic. Theory Moral Pract. **22**(3), 731–747 (2019)
28. De Jong, R.: The retribution-gap and responsibility-loci related to robots and automated technologies: a reply to Nyholm. Sci. Eng. Ethics **26**(2), 727–735 (2020)
29. Adamson, P.: Philosophy in the Hellenistic and Roman Worlds, vol. 1. Oxford University Press, Oxford (2015)
30. Kant, I.: Groundwork for the Metaphysics of Morals. Oxford University Press, Oxford (1785/2002)
31. Waldron, J.: Dignity, Rank, and Rights. Oxford University Press, Oxford (2012)
32. Nussbaum, M.: Frontiers of Justice: Disability, Nationality, Species Membership. Harvard University Press, Cambridge (2006)
33. Pettit, P.: Just Freedom. Norton, New York (2014)
34. Anderson, E.: Private Government. Princeton University Press, Princeton (2017)
35. Nyholm, S.: Humans and Robots: Ethics, Agency, and Anthropomorphism. Rowman & Littlefield International, London (2020)
36. Swanepoel, D.: Does artificial intelligence have agency? In: Robert, W.C., Klaus, G., Inês, H. (eds.) The Mind-Technology Problem, pp. 83–104. Springer, Berlin (2021)
37. Strasser, A.: Social cognition and artificial agents. In: Müller, V. (ed.) Philosophy and the Theory of Artificial Intelligence, pp. 106–117. Springer, Berlin (2017)
38. Vanzura, M.: What is it like to be a done operator? Or, Remotelz extended minds in war. In: Clowes, R.W., Gärtner, K., Hipólito, I. (eds.) The Mind-Technology Problem, pp. 211–229. Springer, Berlin (2021)
39. Popa, E.: Human goals are constitutive of agency in artificial intelligence (AI). Philos. Technol. (2021). https://doi.org/10.1007/s13347-021-00483-2
40. Clarke, A., Chalmers, D.: The extended mind. Analysis **58**(1), 7–19 (1998)
41. Vold, K.: The parity argument for extended consciousness. J. Conscious. Stud. **22**(18), 16–33 (2015)
42. List, C.: Group agency and artificial intelligence. Philos. Technol. (2021). https://doi.org/10.1007/s13347-021-00454-7
43. Coeckelberg, M.: Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf. Technol. **12**(3), 209–221 (2010)
44. Gunkel, D.: Robot Rights. MIT Press, Cambridge (2018)
45. Danaher, J.: Welcoming robots into the moral circle: a defence of ethical behaviourism. Sci. Eng. Ethics **26**(4), 2023–2049 (2020)
46. Loh, J.: Roboterethik: Eine Einführung. Suhrkamp, Frankfurt (2019)
47. Schwitzgebel, E., Garza, M.: Designing AI with rights, consciousness, self-respect, and freedom. In: Matthew Liao, S. (ed.) Ethics of Artificial Intelligence, pp. 480–505. Oxford University Press, Oxford (2020)
48. Schwitzgebel, E., Garza, M.: A defense of the rights of artificial intelligences. Midwest Stud. Philos. **39**(1), 98–119 (2015)
49. Wareham, C.S.: Artificial intelligence and African conceptions of personhood. Ethics Inf. Technol. **23**(2), 127–136 (2020)
50. Sneddon, A.: Symbolic value. J. Value Inquiry **50**(2), 395–413 (2016)
51. Richardson, K.: The asymmetrical 'relationship': parallels between prostitution and the development of sex robots. SIGCAS Comput. Soc. **45**(3), 290–293 (2015)
52. Sparrow, R.: Robots, rape, and representation. Int. J. Soc. Robot. **9**(4), 465–477 (2017)
53. Nyholm, S., Frank, L.: It loves me, it loves me not: is it morally problematic to design sex robots that appear to love their owners? Techne **23**(3), 402–424 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.