Chapter 9
AI, Control and Unintended Consequences: The Need for Meta-Values

Ibo van de Poel

Abstract Due to their self-learning and evolutionary character, AI (Artificial Intelligence) systems are more prone to unintended consequences and more difficult to control than traditional sociotechnical systems. To deal with this, machine ethicists have proposed to build moral (reasoning) capacities into AI systems by designing artificial moral agents. I argue that this may well lead to more, rather than less, unintended consequences and may decrease, rather than increase, human control over such systems. Instead, I suggest, we should bring AI systems under meaningful human control by formulating a number of meta-values for their evolution. Amongst others, this requires responsible experimentation with AI systems, which may neither guarantee full control nor the prevention of all undesirable consequences, but nevertheless ensures that AI systems, and their evolution, do not get out of control.

Keywords Artificial intelligence · Control · Unintended consequences · Values · Machine ethics · Value sensitive design · Experimentation · Machine learning

9.1 Introduction

The worry that technology can get out of control is an old one (e.g., Winner, 1977). It has been expressed in stories and cautionary tales like that of Frankenstein and Prometheus, which express the concern that we, as humans, can construct a technology that henceforth becomes autonomous and takes over from us, or has otherwise destructive consequences. It is therefore not amazing that this worry has also been voiced with regards to Artificial Intelligence (AI) (e.g., Bostrom, 2016;

I. van de Poel (✉)
Department of Values Technology and Innovation, School of Technology, Policy and Management, TU Delft, Delft, The Netherlands
e-mail: I.R.vandePoel@tudelft.nl

A. Fritzsche, A. Santa-María (eds.), *Rethinking Technology and Engineering*, Philosophy of Engineering and Technology 45, https://doi.org/10.1007/978-3-031-25233-4_9

Cellan-Jones, 2014). One of the more specific forms that this worry has taken is that AI systems become so (generally) intelligent that they surpass humans and will take over or will eradicate humans as an inferior form of intelligence.

Although this is an intriguing worry, it does not seem a very realistic one, at least not in the foreseeable future. We, for example, seem to have currently more reason to worry about AI systems that are not intelligent enough for the tasks we let them carry out, than about AI systems that become too intelligent (Levesque, 2017). Nevertheless, there seem to be good reasons to worry that AI systems can autonomously evolve in undesirable ways due to their adaptive characteristics. It is, for example, well conceivable that AI systems will disembody the values for which they were initially designed (cf. Vanderelst & Winfield, 2018; Cave et al., 2019).[1]

In this chapter I discuss how we can keep AI systems under human control. To do so, I start with exploring the specific characteristics of AI systems, such as autonomy, interactivity and adaptability, that distinguish them from more traditional sociotechnical systems. I argue that these characteristics make it more likely, and harder to avoid, that AI systems will have unintended consequences. To deal with such unintended consequences, I consider two proposed approaches, designing AI for human values and machine ethics, and argue that these both fall short by insufficiently addressing the evolutionary character of AI. I then suggest that in order to bring AI systems under meaningful human control, we need a set of what I call meta-values, such as transparency, accountability and reversibility, that apply to the *evolution* of AI systems. The consequent approach treats the introduction of AI systems in society, and their subsequent evolution, as a moral experiment, and accepts that while we will not be able to anticipate all consequences of their employment, nor can avoid all unintended consequences, we nevertheless should ensure that evolutionary AI systems remain correctable and leave enough room for human intervention.

9.2 What Is AI and What (If Anything) Is Special About It?

There are many definitions of AI. Here I employ a broad definition or characterization along the following lines: AI systems are systems that can carry out tasks that, if carried out by humans, would require intelligence. This definition does not say that AI systems themselves are intelligent, whatever that would exactly mean.[2] It also does not assume a specific set of techniques that need to be employed to call something AI. I take AI here to cover the broad field of so-called Good

[1] We may say that an AI systems disembodies a value V if it adapts itself in such a way that it is no longer conducive to V (under normal circumstances), even if it originally embodied V (Van de Poel, 2020).

[2] It indeed also does not say what intelligence is. That is, of course, a huge philosophical question. For now, I am just assuming that we have at least a rough idea of what tasks require human intelligence and which ones not.

Old-Fashioned AI (GOFAI), connectionism, as well as dynamic approaches (Walmsley, 2012). GOFAI goes back to 1950s and 1960s and is based on a representational theory of (human) cognition and the mind; it can be understood as trying to build a computer model of the mind, based on the idea that the mind largely functions as a representational device. Connectionism employs neural networks, and might be said to be based not on a representational but on a neurological (or brain) model of the (human) mind. Many of the current machine learning (ML) techniques are based on neural networks (e.g., Russell & Norvig, 2016). Dynamic approaches are based on an embedded notion of cognition, where cognition is not just in the mind but also in the (human) body and the environment, also sometimes expressed in terms of the 'extended mind' (Clark & Chalmers, 1998). Many approaches in robotics seem to be consonant with this idea (e.g., Bruno et al., 2018).

Although the focus is often on specific AI techniques, algorithms or applications, AI systems are probably best conceptualized as sociotechnical systems.[3] Sociotechnical systems are systems that consist of three basic building blocks, namely technologies (or technological artifacts broadly conceived), human agents and institutional rules (Ottens et al., 2006). The latter refers to the social rules that regulate the behavior of human agents vis-à-vis each other, and vis-à-vis technologies. Elsewhere, I have argued that AI systems consist of two additional building blocks compared to traditional sociotechnical systems, namely artificial agents and what I call 'technical rules' (Van de Poel, 2020). Artificial agents can carry out human-like tasks and roles in a sociotechnical system. The term 'technical rules' may be somewhat confusing in this context as it does not refer to the rules humans (should) follow in designing or acting with technology, rather it is meant here as an equivalent to social rules, but in this case regulating the behavior of artificial agents vis-à-vis each other, and vis-à-vis other elements of the systems (viz., technological artifacts and humans).

Given this broad characterization of AI, one might wonder whether there is anything that AI systems have in common and that sets them apart from other sociotechnical systems. I believe the crux is to be found in the fact that AI systems also contain artificial agents which may have properties such as 'autonomy', 'interactivity' and 'adaptivity' (cf. Floridi & Sanders, 2004), properties that traditionally only human agents have in a sociotechnical system. I take artificial agents to be *autonomous* in the sense that they have the capacity to adapt their own behavior or mode of operation without interference from the environment, and more specifically without human interference. Artificial agents are not just autonomous, they are also *interactive*, which means that they (can) interact with their environment, both in the sense that they act upon their environment, and can hence affect or change it, as well as in the sense that they can pick up signals from their environment. Combined with their autonomy, this interactivity makes artificial agents adaptive, in the sense that

[3] It is therefore somewhat unfortunate that much of the discussion about the ethical and social implications of AI has focused on algorithms, while many of the concerns are raised, and need to be addressed, at the level of AI systems as sociotechnical systems.

they can pick up signals from the environment and autonomously adapt their own functioning on basis of these signals.

I take the inclusion of artificial agents with autonomy, interactivity and adaptability to be characteristic for AI systems, and to set them apart from other socio-technical or engineering systems.[4] Therefore, in the remainder of this chapter I will focus on what the combination of these three characteristics implies for the way in which AI systems have unintended consequences, and for how we can, or cannot, control these consequences and the (moral) values that should be upheld in the development of AI systems.

9.3 Unintended Consequences

Like any other technology, the use of AI may have unintended consequences. It is useful to distinguish between three broad causes of such unintended consequences, namely:

(1) **Lack of due care** in the development and employment of the technology for unintended consequences, in particular lack of foresight or anticipation of the potential effects of the employment of a technology;
(2) **Epistemic ignorance**, i.e., lack of knowledge. Here I am particularly interested in those cases where this lack of knowledge is not the result of a lack of due care (category 1), but rather is of a more fundamental nature, i.e., those cases in which developers and users of a technology could not have reasonably foreseen the unintended consequences;
(3) **Indeterminacy**, i.e., situations in which the causal chain towards the ultimate (unintended) consequences is still open, and in which their occurrence for an important part is determined by agents or factors beyond the control of those designing (and employing) the technology.

Whereas in the second case, unintended consequences may be hard or impossible to prevent due to a lack of knowledge, and hence to limitations in our ability to know certain things, in the third case, the consequences of the employment of a technology are underdetermined in a more ontological sense because the possibility of unintended consequences depends on things that still need to happen or choices that are still to be made.

In a concrete situation, the distinction between the three categories may be sometimes hard to make. For example, we may not always know whether we are in a situation of indeterminacy or of epistemic ignorance (cf. Poser, 2013). Similarly, once certain unintended consequences have materialized, it may be debatable whether

[4] In current (ethical) discussions about AI systems, opacity is also often seen as a typical characteristic of AI systems. While AI systems may indeed be opaque, and this may raise ethical worries, it is in my approach here not a characteristic of *all* AI systems, as it very much depend on the AI techniques employed; in general ML techniques are much more prone to opacity than GOFAI.

these could have been prevented if due care had been exercised or are due to a more fundamental form of epistemic ignorance. Nevertheless, the distinctions are useful to compare the potential causes of unintended consequences between different technologies. This allows us to say something about which technologies are more prone to unintended consequences that are hard if not impossible to foresee and prevent.

Here the three characteristics of artificial agents – autonomy, interactivity and adaptivity – are particularly relevant. It would seem that these characteristics make AI systems more susceptible for indeterminacy compared to traditional sociotechnical or engineering systems. After all, these characteristics mean that the properties of an actual AI system do not just depend on how the system has been designed, and how it is (currently) used, but also on how it has evolved over time, and will evolve in the future. While this is also true for traditional sociotechnical systems, AI systems seem to have a higher likeliness to evolve in indeterminate ways, particularly in ways not intended or foreseen by the human agents in the system. The outcome of this evolution is not only indeterminate but it may also very hard to know how the system would possibly evolve (i.e., epistemic ignorance).

More than traditional engineering systems, AI systems then are plagued by indeterminacy and epistemic ignorance when it comes to the occurrence of unintended consequences. This has not only consequences for how likely it is that such intended consequences occur, but also for the effectiveness of current strategies to avoid possible unintended consequences. I will first discuss in the Sects. 9.4 and 9.5 existing approaches to deal with unintended consequences, and then propose a somewhat new approach in Sect. 9.6 and later. This new approach builds on design for values approaches discussed in Sect. 9.4 but extends it with a new set of meta-values for the evolution of AI systems (Sect. 9.6), as well as a more 'experimental' approach (Sects. 9.7 and 9.8).

9.4 Designing AI Systems for Human Values

One of the existing approaches for avoiding undesirable unintended consequences of AI is to design such systems pro-actively for human values. There is now an extensive literature available on AI ethics and a range of human values and moral principles have been articled that should be adhered to in the design of AI systems. For example, the European High-Level Expert Group on AI (2019) has articulated the ethical principles of respect for human autonomy, prevention of harm, fairness and explicability.

There is indeed much to be said for designing AI systems for such human values and moral principles. However, such strategies basically amount to increasing the due care for unintended consequences in the design of AI systems. This is likely to reduce unintended consequences, and it may also diminish some of the current moral pitfalls in the employment of AI systems, such as bias and opacity. However, it will most likely not eliminate the occurrence of unintended consequences due to epistemic ignorance and indeterminacy, and, as we have seen, these factors are

larger for AI systems than for most traditional sociotechnical systems. So while designing AI for human values is necessary to address potential unintended consequences, it will not be enough.

One way of stating this issue is to say that most of the (moral) values and principles that have until now been articulated for the design of AI systems tend to address such systems at the object level. That is to say they are aimed at embedding certain values in the design of an AI system upfront, but they do, as such, not address the further evolution of these systems. However, AI systems often get their shape due to how they evolve and adapt themselves in interaction with their environment rather than due to their initial design. Some have therefore argued that we need to design AI systems that do not just meet a range of human values, but also have themselves reasoning capacities so that they keep their own employment and evolution within certain moral boundaries (e.g., Wallach & Allen, 2009).

9.5 Machine Ethics

One proposed approach to deal with the evolutionary character of AI systems, and the fact that such systems are more prone to unintended consequences is to try to build AI systems that have the capability to keep their own development within certain moral boundaries. Some believe that this requires AI systems with certain moral capabilities (e.g., Wallach & Allen, 2009). This indeed is the basic idea behind what is often called machine ethics (e.g., Anderson & Anderson, 2011).

Machine ethics may be seen as an attempt to reduce the unintended consequences of AI and to keep these systems within certain moral boundaries by building moral capabilities into AI systems themselves. While this motivation is certainly laudable, the offered solution seems me mistaken for at least two reasons.[5] First, we are still very far removed from anything like artificial moral intelligence (Winfield, 2019; Müller, 2020). There are many reasons for this, one of them being that we do not yet fully understand what grounds or makes up human moral capabilities. So if it really were true that machine ethics is needed in order to responsibly develop AI systems, we would seem to have good reasons for a moratorium on AI as artificial moral intelligence is really not ready for its task yet. (Luckily, it is not true, and there are other ways to oversee the evolution of AI systems as we will see below).

Second, even if it were possible to develop artificial moral agents with the required moral capabilities, it would very likely not be enough to prevent unintended consequences, as it will not take away the more fundamental reasons why AI systems has unintended consequences that I have discussed before such as epistemic ignorance and indeterminacy. On the contrary, if there is one thing we can learn from the philosophy and history of technology then it is that by developing

[5] For criticism of the various reasons that have been given for developing artificial moral agents, see Van Wynsberghe and Robbins (2019).

complex systems that aim to better control their environment, we may well increase rather than decrease the amount of unintended consequences. One reason is that in such cases, we tend to create more complex and more tightly coupled systems that are more vulnerable to unexpected or unforeseen events (cf. Perrow, 1984; Collingridge, 1992).

9.6 Meaningful Human Control: The Need for Meta-Values

Rather than aiming to let AI control its own development, we better bring it under what has been called meaningful human control. The term 'meaningful human control' was initially coined for the development and use of military drones, as it was considered undesirable – in terms of morality and international law – to develop and use drones that would autonomously decide to kill a potential enemy (Horowitz & Scharre, 2015; UNIDIR, 2014). The idea is that such decisions should be ultimately made by humans in a meaningful way, i.e., based on sufficient and adequate information, with proper time to reflect and decide.[6]

The principle has been developed into a more general one for the design of AI and robot systems (Santoni de Sio & Van den Hoven, 2018). Here I am not so much interested in applying the principle at the object level, like the design of a specific AI or robot system (like a drone or autonomous car), but rather as a principle that should be adhered to in the evolution of AI systems. In that case, it would require that we only allow AI systems to evolve in such a way that the process of their evolution remains under meaningful human control.

What would this exactly mean? First, we would need to put in place ways to monitor how AI systems evolve over time and to intervene in this evolution if necessary. This means that we need to extend the value-sensitive design of such systems to their entire life cycle rather than to restrict it to their initial design (Umbrello & Van de Poel, 2021; De Reuver et al., 2020). AI systems may, over time, require undesirable properties or disembody their initial embedded values. Meaningful human control would therefore mean that we can reverse this process if necessary. This can, for example, be effectuated by letting an AI system make over time stored versions of itself, so that we can return to an earlier version, not unlike what we do with software updates. Meaningful human control would then imply that the

[6] This is not to suggest that in traditional warfare, decisions to kill an enemy are, or can, always be made in a meaningful way. This is often obviously not the case. Meaningful control in this situation seems more like an ideal, and the relevant moral question with respect to introducing autonomous or semi-autonomous drones in war(like) situations seems to be whether they will increase meaningful human control rather than they can fully guarantee it. (There might be of course also other moral considerations that speak for or against the use of drones in warfare or similar situations).

evolution of AI systems should meet some minimal requirements in terms of reversibility and adaptability.[7]

In order to achieve control that is meaningful, we would also need some understanding of how the system has evolved over time. This requires adherence to values similar to ones that are now already often mentioned in relation to AI like explainability, transparency and accountability. Again, however, these are now usually applied at what I have called the object level. For example, if an AI system is to make important decisions, like in a court case, we want those decisions to be transparent and explainable.[8] Here, however, I am interested in the application of such values to the evolution of AI systems. At this level, these values are required because if we believe that an AI system has evolved in an undesirable way, we want to know why and how it did, both to be able to return to an acceptable earlier point in its evolution, as well as to avoid such an undesirable evolution to occur again in the future.

We might call the values that are needed to retain meaningful human control over an AI system during its evolution meta-values. Such meta-values do not apply to AI at the object level, but rather set constraints on how we allow AI systems to evolve over time. If meta-values are to be guaranteed also during the evolution of an AI system, it means that we need to build these values in an immutable way into AI systems, so that they cannot be disembodied during the evolution of the AI systems; these are hence to be designed as hard constraints (or for example technical rules) into the system.

The above discussion suggests some candidates for meta-values, such as reversibility, adaptability, accountability and transparency. There might be more meta-values than these ones. Some might want to argue that also other values often mentioned in AI ethics like non-maleficence (doing no harm) and fairness should be seen as meta – values. After all, it seems highly unlikely that in the future we would morally want AI systems that do harm or are unfair. However, it should be noted that what counts as 'harm' and as 'fair' is much more context-dependent, and also more open to future change than reversibility. For such reasons, we might want to restrict meta-values to those values that are needed to keep AI under meaningful control, while the more substantive values can, and should, be addressed at the object level, where we can better do justice to context and value change over time.

[7] I am not suggesting that we should require that all consequences of (the use of) AI systems are reversible; that would seem unfeasible and unrealistic. Rather I would want to require that the evolution of specific AI systems is made reversible, so that we can go back to an earlier moment in their evolution.

[8] Robbins (2019) complains that explainability, or explicability, is often applied to the AI (or ML) system rather than to the decisions made by such systems. It is the latter that in his view should (at least sometimes) be explainable. I agree for those cases in which we consider AI at what I have called the object level. However, at the evolutionary level, values like explainability or explicability would apply to the evolution of the AI system itself. It should be noted that such explainability at the evolutionary level is not enough to guarantee explainability at the object level. So when designing AI systems that make important decisions, we need explainability at both levels.

9.7 An Experimental Perspective

Keeping the evolution of AI systems under meaningful human control by adherence to a set of meta-values as proposed above will not guarantee that AI systems are free from unintended consequences. After all, meaningful human control as such does not reduce epistemic ignorance or indeterminacy. My proposal then is not aimed at preventing all unintended consequences, which would seems me illusionary any way, but rather at ensuring that the evolution of AI systems remains correctable and, to some extent, reversible if undesirable unintended consequences materialize.

The ultimate aim then is not so much to control the development and evolution of AI systems in the sense of strictly guiding or regulating it, but rather to make sure that such systems do not get out of human control. As I have argued elsewhere, we best think of technological development and the introduction of new technology into society as an experimental process (Van de Poel, 2017). This implies that one cannot fully predict the impacts of AI in society beforehand. It also means that we should be willing to accept some risks and unintended consequences. Questions about the acceptability of new technology are in this perspective best formulated in terms of the acceptability of experimenting with technologies like AI in society, rather than in terms of whether the technology as such is acceptable or not (Van de Poel, 2016). So, in addition to guaranteeing that AI systems keep under meaningful human control, we need rules and institutions that allow responsible experimenting with such systems.

9.8 Human Indeterminacy

The experimental perspective implies that we should accept, whether we like it or not, some degree of indeterminacy. Nevertheless, indeterminacy might seem in principle undesirable as it opens up the possibility for unintended consequences. Designers of engineering systems also often want to reduce indeterminacy as it, in their view, typically increases the chances that a designed system will not function as intended.

The traditional engineering approach to indeterminacy then seems to be to try to design it out, for example by reducing the ways in which a technology could be used, or even by (attempts at) enforcing a particular way in which humans can interact with an engineering system (cf. Fritzsche, 2010). For example, to ensure safe operation, engineers often aim to make designs fool-proof by enforcing a certain way of using a system, so that safety risk are less likely to arise (Bucciarelli, 1985; Van de Poel & Robaey, 2017). An example is the lock-out switch above the rear handle on a chain saw; unless both this lock-out switch and the rear handle are pressed, the chain will not be driven. This mechanism increases safety by enforcing a way of using the chain saw, by having to use both one's hands, that makes it much less likely that users will inadvertently saw off their own hand.

While system designers thus often aim at reducing indeterminacy, the case of AI would seem to be different in important ways. As we have seen before, AI systems are self-learning and this makes them inherently more indeterminate than most traditional sociotechnical systems. This indeterminacy is in fact often seen as an inherent and desirable feature, as it allows an AI system to learn from its environment and to adapt itself.[9] The flip side is, of course, that AI systems are also inherently harder to keep under control than traditional engineering systems, and more likely to lead to unintended consequences.

Still there is also something positive about indeterminacy. I think that even an argument can be made that it is desirable to design systems with at least some degree of indeterminacy. To flesh out this argument, it is useful to distinguish between what might be called 'technical indeterminacy' and 'human indeterminacy'. With 'technical indeterminacy', I mean the indeterminacy that is inherent in a sociotechnical system as technical system, i.e., independent from deliberate human interventions in the system, for example due to how a self-learning algorithm develops itself in interaction with its environment. With human indeterminacy, I mean the degree to which a social technical system is open to (deliberate) human interventions.

While there might be good reasons to decrease human indeterminacy for some sociotechnical systems, as in the case of the chain saw, I would like to suggest that it is often desirable to design systems that have at least some degree of human indeterminacy. There are various reasons for that (cf. Van de Poel & Robaey, 2017). First, human indeterminacy allows to be responsive to new developments and to unintended consequences, as it leaves room to intervene in a sociotechnical system also during its operational phase. Second, human indeterminacy creates rooms for system users to make the system really their own, and to appropriate it. It could be argued that this is desirable for democratic reasons, as it, for example, allows users to have different (value) priorities than the original system designers. Similarly, it creates room for future users to use systems in their own way, and to adapt to changing values in society. Third, apart from the previous considerations, some degree of human indeterminacy would seem required to keep AI systems under meaningful human control. For reasons set out before, AI systems will anyway sometimes develop in unexpected ways and have unintended consequences: meaningful human control can in such cases only be assured by some minimal degree of human indeterminacy in order to allow humans to deliberately intervene.

The above discussion then suggests that we should not just willy-nilly accept indeterminacy as an inevitable bad, but that it might actually also be something

[9] It is open to debate whether the (future) evolution of AI systems is really indeterminate or just (epistemically) unknown. It is, however, at least the last I would argue. If we can know, and hence predict, how an AI system will evolve in the future, we would no longer need to build it as a self-learning system. The advantage of a self-learning system after all is that while we do not now the future (or the environment), we can build something that is able to develop in response to how things evolve (or in response to its environment). In a practical sense, openness to the future then is a key characteristic of AI systems.

good, in particular when it comes to human indeterminacy in contrast to technical indeterminacy. A possible counterargument is that increasing human indeterminacy will, in effect, increase the probability of misuse and of unintended and undesirable consequences. I have two replies to such an argument. First, I think we should indeed carefully consider from case to case what degree of (human) indeterminacy is desirable. I would, for example, not argue against the design of chain saws with lock-out switches as the gain in safety would seem me worth the loss in human indeterminacy in this specific case. Second, even if human indeterminacy may increase the possibilities for misuse and undesirable consequences, this may, at least sometimes, be a price worth paying for the various reasons I have discussed before.

9.9 Conclusions

I have argued that three characteristics – autonomy, interactivity and adaptivity – set AI systems apart from traditional sociotechnical systems. These characteristics make the evolution of AI systems more indeterminate, and therefore harder to control and more likely to lead to unintended consequences.

Addressing these new challenges introduced by AI requires three things. First, it requires designing AI systems for human values and anticipating possible negative unintended effects of their employment in society. While this is now fairly widely recognized, this is a necessary but not yet a sufficient condition. Second, it requires assuring meaningful human control over the *evolution* of AI systems; this requires a set of meta-values, like monitorability, reversibility, adaptability and accountability that are guaranteed as immutable values in the evolution of AI systems. Third it requires, new societal modes and institutions to responsible experiment with AI in society as to gradually learn how to best embed it in society and to adjust its course where necessary.

Finally, I have suggested that we should not aim for developing and employing AI systems in a way that is completely fail-safe. Not only is such a goal unattainable and in that sense illusory, it may well backfire, as it may lead to designing out the very indeterminacies in such systems that we need to keep such systems under meaningful human control.

Acknowledgement This publication is part of the project ValueChange that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 788321. This publication also contributes to the research programme Ethics of Socially Disruptive Technologies, which is funded through the Gravitation programme of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.004.031).

References

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press.

Bostrom, N. (2016). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bruno, L., Astorga, D., Mendoza-Bock, E., Pardo, M., Escobar, E., & Ciria, A. (2018). Embodied cognitive robotics and the learning of sensorimotor schemes. *Adaptive Behavior, 26*(5), 225–238. https://doi.org/10.1177/1059712318780679

Bucciarelli, L. L. (1985). Is idiot proof safe enough? *The International Journal of Applied Philosophy, 2*(4), 49–57.

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and risks of machine ethics. *Proceedings of the IEEE, 107*(3), 562–574. https://doi.org/10.1109/JPROC.2018.2865996

Cellan-Jones, R. (2014). Stephen Hawking warns artificial intelligence could end mankind. *BBC*. Accessed April 26, 2020. https://www.bbc.com/news/technology-30290540

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Collingridge, D. (1992). *The management of scale. Big organizations, big decisions, big mistakes*. Routledge.

De Reuver, M., van Wynsberghe, A., Janssen, M., & Van de Poel, I. (2020). Digital platforms and responsible innovation: Expanding value sensitive design to overcome ontological uncertainty. *Ethics and Information Technology, 22*, 257–267. https://doi.org/10.1007/s10676-020-09537-z

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*(15). https://doi.org/10.3389/frobt.2018.00015

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Fritzsche, A. (2010). Engineering determinacy: The exclusiveness of technology and the presence of the indeterminate. In I. van de Poel & D. Goldberg (Eds.), *Philosophy and engineering: An emerging agenda* (pp. 305–312). Springer Netherlands.

High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI*. EC.

Horowitz, M., & Scharre, P. (2015). *Meaningful human control in weapon systems: A primer*. Center for a New American Security.

Levesque, H. J. (2017). *Common sense, the Turing test, and the quest for real AI*. MIT Press.

Müller, V. C. (2020). Ethics of artificial intelligence and robotics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2020 edition).

Ottens, M., Franssen, M., Kroes, P., & Van de Poel, I. (2006). Modeling engineering systems as socio-technical systems. *International Journal of Critical Infrastructures, 2*, 133–145.

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. Basic Books.

Poser, H. (2013). The ignorance of engineers and how they know it. In D. P. Michelfelder, N. McCarthy, & D. E. Goldberg (Eds.), *Philosophy and engineering: Reflections on practice, principles and process* (pp. 3–14). Springer Netherlands.

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines, 29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach* (3rd ed.). Pearson.

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics,*1, 283–296. https://doi.org/10.1007/s43681-021-00038-3

UNIDIR. (2014). *The weaponization of increasingly autonomous technologies: Considering how meaningful human control might move the discussion forward*. UNIDIR (United Nations Institute for Disarmament Research).

Van de Poel, I. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics, 22*(3), 667–686. https://doi.org/10.1007/s11948-015-9724-3

Van de Poel, I. (2017). Society as a laboratory to experiment with new technologies. In D. M. Bowman, E. Stokes, & A. Rip (Eds.), *Embedding new technologies into society: A regulatory, ethical and societal perspective* (pp. 61–87). Pan Stanford Publishing.

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines, 30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

Van de Poel, I., & Robaey, Z. (2017). Safe-by-design: From safety to responsibility. *NanoEthics, 11*(3), 297–306. https://doi.org/10.1007/s11569-017-0301-x

Van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*(3), 719–735. https://doi.org/10.1007/s11948-018-0030-8

Vanderelst, D., & Winfield, A. (2018). The dark side of ethical robots. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New Orleans, LA, USA.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Walmsley, J. (2012). *Mind and machine*. Palgrave Macmillan.

Winfield, A. F. (2019). Machine ethics: The design and governance of ethical AI and autonomous systems. *Proceedings of the IEEE, 107*(3), 509–517.

Winner, L. (1977). *Autonomous technology. Technics-out-of-control as a theme in political thought*. MIT Press.