Embedding Values in Artificial Intelligence (AI) Systems

van de Poel, Ibo

DOI
10.1007/s11023-020-09537-4

Publication date
2020

Document Version
Final published version

Published in
Minds and Machines

Minds and Machines
https://doi.org/10.1007/s11023-020-09537-4

GENERAL ARTICLE

Embedding Values in Artificial Intelligence (AI) Systems

Ibo van de Poel[1]

Received: 13 April 2020 / Accepted: 18 August 2020

Abstract
Organizations such as the EU High-Level Expert Group on AI and the IEEE have recently formulated ethical principles and (moral) values that should be adhered to in the design and deployment of artificial intelligence (AI). These include respect for autonomy, non-maleficence, fairness, transparency, explainability, and accountability. But how can we ensure and verify that an AI system actually respects these values? To help answer this question, I propose an account for determining when an AI system can be said to embody certain values. This account understands embodied values as the result of design activities intended to embed those values in such systems. AI systems are here understood as a special kind of sociotechnical system that, like traditional sociotechnical systems, are composed of technical artifacts, human agents, and institutions but—in addition—contain artificial agents and certain technical norms that regulate interactions between artificial agents and other elements of the system. The specific challenges and opportunities of embedding values in AI systems are discussed, and some lessons for better embedding values in AI systems are drawn.

Keywords Artificial intelligence · Values · Ethics · Sociotechnical system · Value embedding · Institution · Artificial agent · Norms · Multi-agent system

1 Introduction

Nowadays, a lot of attention is being given to ethical issues, and more broadly to values, in the design and deployment of artificial intelligence (AI). Recently, the EU High-Level Expert Group on AI (2019: 12) formulated four ethical principles that AI applications should meet: respect for human autonomy, prevention of harm, fairness and explicability. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019: 4) also recently formulated a number of high-level principles,

✉ Ibo van de Poel
i.r.vandepoel@tudelft.nl

[1] Department of Values, Technology, & Innovation, School of Technology, Policy & Management, Technical University Delft, Delft, The Netherlands

Published online: 01 September 2020

including human rights, well-being, data agency, transparency, and accountability. These values, as well as relevant others such as security and sustainability, are supposed to guide the governance and design of new AI technologies. But how can we verify or at least assess whether AI systems indeed embody these values?

The question of whether and how technologies embody values is not new. It has been discussed in the philosophy of technology, where several accounts have been developed (e.g., Winner 1980; Floridi and Sanders 2004; Flanagan et al. 2008; Klenk 2020; for an overview of several accounts, see Kroes and Verbeek 2014). Some authors deny that technologies are, or can be, value-laden (e.g., Pitt 2014; for a criticism, see Miller 2020), while others see technologies as imbued with values due to the way they have been designed (e.g., Winner 1980; Flanagan et al. 2008; Van de Poel and Kroes 2014). Still others treat technologies as moral agents, somewhat similar to human agents (e.g., Floridi and Sanders 2004; Sullins 2006; Verbeek 2011), and some even argue for abandoning the distinction between (human) subjects and (technological) objects altogether in understanding how technologies may embody values (Latour 1992, 1993).

An account of value embodiment in technology can help in assessing whether designed AI systems indeed embody a range of moral values, such as those articulated by the EU High-Level Expert Group and the IEEE. For such an account, three desiderata stand out. First, the account should be connected to the design of AI systems, as this appears to be an important target in ethical codes for AI (e.g., IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019). Second, the account should treat AI systems not just as isolated technical artifacts but as sociotechnical systems (Borenstein et al. 2019; Coeckelbergh 2020; Boddington 2017; Behymer and Flach 2016; Jones et al. 2013). Moreover, it should fully consider the fact that AI systems are in some respects different from traditional sociotechnical systems. Third, the account should be able to maintain at least some (conceptual) distinctions between humans and AI and therefore not treat AI systems as moral agents similar to human agents (cf. Johnson 2006; Johnson and Miller 2008; Illies and Meijers 2009; Peterson and Spahn 2011). This is important not just for philosophical conceptual reasons but also for moral reasons. If we want to do justice to a value like "respect for human autonomy," we need a conceptual framework that distinguishes human (or moral) autonomy and agency from the potential autonomy and agency of AI.

To meet these desiderata, I build on the account presented in Van de Poel and Kroes (2014). This account meets the first and third conditions. It meets the first condition by accounting for embodied values (in a technology) as the result of certain intentional value-embedding activities by designers. In other words, it holds that under certain conditions, designers can successfully embed values in a technology, which may then be said to embody those values. It meets the third condition by treating technologies as value-laden without conceiving of them as moral agents or otherwise diluting the difference between humans and technologies.

As it stands, the account in Van de Poel and Kroes (2014) does not yet meet the second condition, because it focuses on technical artifacts rather than on sociotechnical systems or AI systems. Such systems are typically hybrid in nature, that is, they contain human as well as technological components. Traditional sociotechnical

systems consist of three basic building blocks: technological artifacts, human agents, and institutional rules (Kroes et al. 2006; Franssen 2015; Ottens et al. 2006). What sets AI apart from traditional technologies is its capacity to autonomously interact with its environment and to adapt itself on the basis of such interactions. This capacity creates new opportunities for embedding values in AI systems that do not exist in traditional sociotechnical systems (cf. Wallach and Allen 2009; Anderson and Anderson 2011). At the same time, however, the adaptivity of AI may undermine the embodiment of values, as it may result in the—perhaps unintended—disembodying of values that were originally embedded by the system designers (cf. Cave et al. 2019; Vanderelst and Winfield 2018).

To address the specific character of AI, this article understands AI systems as consisting of additional building blocks, beyond technological artifacts, human agents, and institutions. The first and main additional building block is artificial agents (AAs). Like traditional technical artifacts, AAs are designed and can embody values, but unlike technical artifacts, they are autonomous, interactive, and adaptive (Floridi and Sanders 2004). Due to these properties, the way they embody values is different from how technical artifacts embody values, particularly because AAs may also contain representations of values (Moor 2006). While AAs may play roles similar to those played by human agents in AI systems, they differ from the latter in the sense that they do not have human intentions and moral agency. In addition to AAs, AI systems contain a fifth building block, here called technical norms. Whereas in traditional sociotechnical systems, institutions regulate the interactions between human agents (and their interactions with technical artifacts), for AAs, this role is played by so-called technical norms. While technical norms may be represented in a syntax and semantics similar to those of institutions, their functioning does not ultimately rest on human intentions, as is the case with institutions, but on the (causal) laws of nature.

This article aims to extend Van de Poel and Kroes's (2014) value-embedding account to AI systems. To do so, it first offers a conceptualization of values and of what it means to embed values in technology. Next, it provides a conceptualization of sociotechnical systems and what distinguishes traditional sociotechnical systems from AI systems. This results in five building blocks that are considered the main components of AI systems: technical artifacts, institutions, human agents, artificial agents, and technical norms. For each of these building blocks, I discuss whether and how it can embody value. This discussion of value-embedding in the five basic building blocks then culminates in a discussion of when an AI system as a whole can be said to embody certain values. In conclusion, I propose a few tentative lessons for the better embedding of values in AI systems.

2 What are Values?

Defining "value" is notoriously difficult as the notion is used widely, not only in daily language but also in various disciplines such as philosophy, economics, sociology, psychology, and anthropology (e.g., Brosch et al. 2016; Hirose and Olson 2015). Nonetheless, "value" is typically associated with what is "good" or

"desirable." Rather than being descriptive, values are normative and express what is "good." More precisely, values can be situated in the evaluative part of normativity, which is distinguished from the deontic part of normativity. Values and other evaluative notions are used to evaluate states of affairs or other entities such as technological artifacts in terms of goodness and badness. Conversely, deontic notions, such as duties, norms, and reasons, are used to determine the rightness (or wrongness) of actions.

Sometimes value is understood as the result of valuing (e.g., Stevenson 1944). Consequently, values may be understood as that which people value. The problem with such an understanding is that people might very well value things that are not valuable; they sometimes even value things that they know they should not value. Conversely, people might sometimes fail to value things that are valuable.

To avoid these problems, values should be understood in relation to normative reasons (cf. Scanlon 1998; Raz 1999; Zimmerman 2015; Jacobson 2011; Anderson 1993). The idea is that there exists a certain correspondence between normative reasons for valuing and for something being of value. Thus, if something is of value, there exist normative reasons to value it, but that does not mean it is also always actually valued, as people may fail to value on the basis of normative reasons. Conversely, this also does not mean that if something is valued, it is of value, as people sometimes value on the basis of wrong reasons or no (normative) reasons at all.

This account of value is helpful in interpreting what it means to say that some entity—in this case, a technical artifact or a sociotechnical system—embodies a value. For an entity to embody a value, there must be reasons for a pro-attitude or pro-behavior toward that entity. For example, if a painting is beautiful, we have reason to admire it. Similarly, if a technical artifact embodies a value, we may have reason to use it or to use it in a particular way (that respects the relevant value).

However, the mere presence of reasons for a pro-attitude or pro-behavior does not show that an entity embodies a certain value; those reasons for a pro-attitude or pro-behavior need to originate in the entity itself and not in something else. This problem is known as the wrong kind of reasons problem (Jacobson 2011). For example, if I promise to give you a certain object, that promise corresponds to reasons for a pro-behavior toward that object (e.g., to protect it against theft), but these reasons originate in my promise, not in the object. Reasons that originate outside the object itself are called the wrong kind of reasons, as they situate the value outside the object (e.g., in the promise made or in the agent making the promise) rather than in the object itself. To avoid the wrong kind of reasons problem, a more detailed account is needed of when an entity, such as a technical artifact, embodies a value.

3 Embodied Values

To assess whether AI systems comply with the values that have been articulated in various codes of ethics for AI, we need further specification of what is meant by such compliance. One understanding of compliance states that the designers of AI systems should be led by those values and should aim to integrate them into the systems they design: so the explicated ethical values should align with the *intended*

values of the system designers. However, intended values would seem too small a basis to assess compliance with ethical values because an intended value can be present even if the designed system fails to fulfill that value.

Another understanding of compliance therefore focuses on the values that are actually realized in the operation of an AI system. A focus on such *realized values*, however, also has drawbacks. First, realized values can only be known once the system is in operation; ideally, one would want to be able to assess a designed system's compliance with moral values before it is actually employed. Second, not all realized values can be meaningfully attributed to the relevant AI system. For example, suppose a self-driving car (understood here as an AI system) causes an accident resulting in a number of fatalities. Can we conclude from this accident that the AI system (i.e., the self-driving car) was unsafe because that value was realized in the accident? The answer seems negative. One accident may not be enough to call a system unsafe. Moreover, the accident may have resulted from exceptional circumstances or irregular use and hence may not be inherent in the system.

The underlying problem is that both intended values and realized values are vulnerable to the wrong kind of reasons problem. In the case of intended values, reasons for a pro-attitude (or con-attitude in the case of disvalue) are grounded in the designers' intentions (and underlying values) rather than in the designed AI system; in the case of realized values, the reasons may be grounded either in the (mis)use of the system or in an unfortunate (but exceptional) outcome rather than in the system itself. If we want reasons that are grounded in the designed AI system itself, we should focus on *embodied value* rather than on intended or realized value.

But how can we understand embodied value in the case of AI systems? The basic idea, which will be further explored and detailed below, is that embodied values should be understood as values that have been intentionally, and successfully, embedded in an AI system by its designers. For a value to be successfully embedded by a designer, two types of conditions need to apply. First, the system has to be intentionally designed to comply with that value. Second, the system has to actually respect or further that value when it is used properly. This idea aids understanding of the relation between intended, embodied, and realized values (see Fig. 1).

The intended values are the values intended by the system's designers. However, these intended values may be different from the embodied values when an artifact (or institution or system) has not been properly designed. The embodied value is the value that is both intended (by the designers) and realized if the artifact or system is properly used. The realized value, in turn, may be different from the embodied value: for example, because a technology is used differently than intended or foreseen. The differences between embodied, intended, and realized values may give raise to different kinds of feedback loops or iterations.

For example, if the realized value is different from the intended value, three types of feedback loops may be activated. First, in situations in which the intended and embodied values are the same, one may try to change the use of the system without necessarily changing its design. However, if the embodied value is different from the intended value, a change in use will not suffice; a change in the design will be required as well. Third, there may also be a category of cases in which the realized values are due to unintended (and unforeseen) consequences. In such cases,

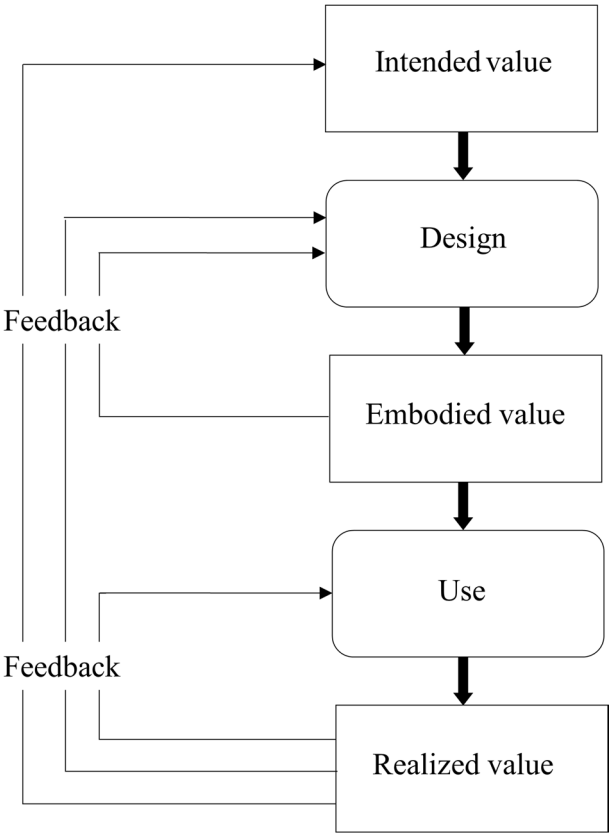

Fig. 1 The relation between intended, embodied, and realized values (adapted from Fig. 7.1 in Van de Poel and Kroes 2014)

a revision of the intended values may be required to avoid the unintended consequences (in the future). For example, if it unexpectedly turns out that an AI system leads to discrimination, it may be necessary to include a new value, such as fair treatment or freedom from bias, among the intended values that designers embed in the system.

The feedback loops in Fig. 1 underscore that design is not a one-off activity. It is in fact an ongoing activity after a system is operational. Such design activities that relate to an already existing and operational system may be called *redesign*. They often do not address the entire system but only parts of it. They may be undertaken not only by professional designers and engineers but also by users and system operators. Redesign is important in any sociotechnical system in which the dynamics of the system is beyond the control of the original system designers, but it is particularly important in the case of AI systems, which due to the adaptive abilities of AI, may acquire system properties that were never intended or foreseen by the original designers.

Table 1 The basic building blocks of an AI system

	Intentional	Physical-Causal
Artifacts	Technical artifacts	
Agents	Human agents	Artificial agents
Rules	Institutions	Technical norms

4 AI Systems as Sociotechnical Systems

In the literature, various definitions and characterizations of sociotechnical systems have been proposed (e.g., Bauer and Herder 2009; Baxter and Sommerville 2011; Geels 2004; Bruijn and Herder 2009; Pasmore and Sherwood 1978; Kroes et al. 2006; Ottens et al. 2006; Dam et al. 2013; Franssen 2014; Nickel 2013). Sociotechnical systems will here be understood as systems that depend on not only technical hardware but also human behavior and social institutions for their proper functioning (cf. Kroes et al. 2006). According to this understanding, sociotechnical systems consist of combinations of three basic building blocks:

1. Technical artifacts
2. Human agents
3. Institutions (rules to be followed by the agents).

What sets AI systems apart from other sociotechnical systems is that they also contain an artificial variety of building blocks two and three. These will be called "artificial agents" and "technical norms," respectively. Whereas the human/societal variety of agents and rules is typically understood in intentional terms, artificial agents and technical norms are ultimately understood in causal or physical terms (see Table 1).

The first building block of sociotechnical systems is technical artifacts. According to the dual nature account developed in the philosophy of technology (e.g., Kroes 2010; Kroes and Meijers 2006), technical artifacts are physical objects that (can) fulfill, and have been designed for, a certain technical function. To understand this function requires not only descriptions in terms of the laws of nature but also references to (human) intentions. Technical artifacts thus have a physical as well as an intentional nature. Following the use plan characterization of technical artifacts developed by Houkes and Vermaas (Houkes et al. 2002; Houkes and Vermaas 2010; Vermaas and Houkes 2006), technical artifacts may be more precisely understood as combinations of physical structures and use plans. A use plan is a plan that describes how an artifact should be used to achieve certain goals or to fulfill its function. In other words, a use plan describes the proper use of a technical artifact, and that proper use will result (in the right context and with users with the right competences) in the artifact fulfilling its proper function.

The second building block is agents, which can be either human agents or artificial agents. The presence of AAs is what sets AI systems apart from other more traditional sociotechnical systems that consist of technical artifacts, institutions, and

human agents, but no AAs. The underlying idea here is that AAs have properties that distinguish them from (traditional) technical artifacts. This includes properties such as autonomy, interactivity, and adaptivity (Floridi and Sanders 2004) that are also possessed by human agents. Floridi and Sanders (2004) therefore suggest that such artificial agents can also become moral agents, but that terminology is misleading (cf. Johnson and Miller 2008) because moral agency, as a philosophical notion, is restricted to agents that possess characteristics such as intentionality, free will, and consciousness.

Which human-like properties (and skills) may be designed into or acquired by AAs is open to debate. Still, it seems likely that there are at least some human characteristics that AAs can never acquire, or at least not in the foreseeable future. These include characteristics such as consciousness, free will, emotions, intentionality, moral autonomy, and moral agency. Although any specific list of what characteristics distinguish AAs from technical artifacts, on the one hand, and from human agents, on the other, is bound to be somewhat controversial, the basic underlying idea that they can be distinguished from both technical artifacts and human agents seems plausible. This would justify treating them as a separate building block in AI systems.

The third building block of sociotechnical systems is rules. In the case of human agents, such rules can be understood as institutions. Institutions are usually conceptualized as specific kinds of *social* norms or rules (e.g., North 1990; Calvert 1995; Ullmann-Margalit 1977; Ostrom 2005; Bicchieri 2006). Such rules prescribe to (human) agents how to behave in a certain (kind of) situation and are typically based on shared expectations, which may be upheld by sanctions if the rule is not followed. Although human agents are able to deviate from the rule, it typically comes at a certain price. Still, not all institutions come with explicit sanctions; some may simply be based on the expectation that other agents will behave similarly.

Institutions are *social* constructs and hence cannot be (directly) perceived and followed by artificial agents. Nevertheless, most artificial multi-agent systems contain an equivalent to institutions, which are here described as *technical norms* (e.g., Mahmoud et al. 2014). The word "technical" is used here not so much because of the content of these rules but because their functioning is ultimately to be understood in causal-physical terms rather than in intentional terms. After all, artificial agents do not have intentions, even if their behavior may sometimes seem intentional.[1]

Now that we have an overview of the main components of a sociotechnical system, we can turn to the question of when a sociotechnical system can be said to embody certain values. To address that question, I first discuss how each of the individual building blocks (technical artifacts, institutions, human agents, artificial agents, and technical norms) may embody values before discussing whether a sociotechnical system as a whole may be said to embody certain values.

[1] This is, of course, not uncontroversial. For a strong defense of the view that AAs, and more generally AI, cannot have intentionality, see Searle (1984). For an opposite view, see, e.g., Dennett (1987).

5 How Technological Artifacts Embody Values

According to Van de Poel and Kroes (2014), technical artifact x embodies value V:

> If the designed properties of x have the potential to achieve or contribute to V (under appropriate circumstances) due to the fact that x has been designed for V.

Hence, two conditions must be met for technological artifact x to embody value V:

1. x is designed for V
2. (The use of) x is conducive to V.

Moreover, these two conditions need to be connected: (the use of) x is conducive to V because x has been designed for V.

The following examples illustrate this account:

1. Sea dikes have been designed to protect against flooding (which may be seen as the proper function of a sea dike). This means that they have been designed for the value of safety (against flooding), and they are conducive for protection against flooding (it is assumed). Given the proposed conditions, they therefore embody the value of safety.
2. A bread knife has been designed to cut bread. It can, however, also be used for killing, and in that sense, it may be conducive to killing. However, since it has not been designed for killing (and killing is not part of its use plan), it does not embody the (dis)value of killing, because it does not meet the first condition.
3. A third example is a badly designed pacemaker. Such a pacemaker presumably meets the first condition in that it has been designed for (contributing to) human well-being. However, it does not embody that value, because it does not meet the second condition: due to the bad design, it fails to contribute to well-being.

These examples illustrate how, in some cases, the account concludes that a technological artifact embodies a value, while in others, it concludes that a technical artifact does not embody a value because either that value was not intended by the designers or the artifact is not conducive to that value.

While the account fares reasonably well in these first three types of cases, there is a fourth type in which the account seems more questionable. These are cases in which unintended consequences systematically occur, which would seem to suggest that a certain value is embodied in a technical artifact, even if that result was never intended by its designers. Consider the following case:

4. A recommender system designed to serve its customers may unintendedly (and systematically) contribute to filter bubbles and echo chambers and thus contribute to (dis)values such as lack of respect and untruth, although that was never intended by its designers.

Due to the lack of design intentions, the account would say that the (dis)values of disrespect and untruth are not embodied in the recommender system. This might seem a mistaken conclusion. One possible solution may be to give up the reference to intentions altogether (cf. Klenk 2020), but that would force us to say that in cases like the bread knife example (example 2), a disvalue is embodied. This conclusion is undesirable because there are always possibilities for misuse or alternative use, and if each (dis)value that could be realized through such unintended use were to be seen as embodied in a technical artifact, technical artifacts would simply embody too many values.

Therefore, another way to address cases of unintended consequences is not to lift the intentionality condition but to become aware that technical artifacts are typically redesigned after their initial design and that such redesign may embed new values in the technical artifact. According to the earlier mentioned use-plan account of technical artifacts, not only designers but also users may redesign a technical artifact, for example by using it in another way than originally intended. Vermaas and Houkes (2006) describe three ways in which artifacts may be used: passive using, idiosyncratic using, and innovative using. Passive using is basically use according to the (original) use plan communicated by the (original) designers. Idiosyncratic use is use that deviates from the use plan in order to achieve certain user goals but without any communication to other prospective users. Innovative using involves the development of a new use plan and communication of it to other users. Innovative using may be interpreted as a form of redesign, as it involves the (successful) design of a new use plan, even if the physical structure of the technical artifact does not change.[2] If we conceive of a technical artifact as a combination of a physical structure and a use plan, technical artifacts with the same physical structure but different use plans may embody different values.[3]

Technical artifacts may therefore begin to embody new values when they are redesigned. Thus, values that are realized but unintended may become realized and intended, and thus embodied. That seems particularly likely to happen in the case of unintended but desirable consequences. But what if the unintended consequences are undesirable, as in example four? Here, the important point is that when certain disvalues are systematically realized, the designers (and users) eventually acquire the obligation to redesign the artifact to avoid those disvalues.

The problem with unintended consequences is that usually they "are not *not* intended" (Winner 1977: 97). But when negative consequences systematically occur, it is at some point no longer good enough to not intend disvalue; an obligation arises to avoid disvalue, and hence to intend, and design for, the opposite positive value. So, in cases like example 4, one may say that the technology fails to embody certain positive values rather than saying that it embodies disvalues.

[2] Vermaas and Houkes (2006) also distinguish what they call expert redesigning in which experts redesign the use plan or provide (scientific) explanations for why a newly developed use plan works. One may assume that in many cases, innovative using and expert redesigning will, when successful, eventually also result in a new design of the physical structure, or what Vermaas and Houkes (2006) call product designing.

[3] Postphenomenologists call this "multistability" (Ihde 2012).

6 Values Embodied in Institutions

The above account can also be applied to institutions. Institutions can be understood as rules that can either be formal (like legal rules or operational instructions) or informal (North 1990; Ostrom et al. 1994). The ADICO grammar developed by Crawford and Ostrom (1995) provides a basis for analyzing institutional rules. This grammar helps to distinguish between different kinds of institutions, and it also helps to identify the basic elements necessary to speak about an institution.

The ADICO grammar analyzes institutions in terms of the following:

A: attributes—to whom a particular institution applies
D: deontic operator—permission (may), obligation (must), or prohibition (must not)
I: aim—actions or results to which the deontic operator applies
C: conditions—describe when, where, how, and to what extent the deontic operator applies
O: or else—describes the sanctions for not observing an institution.

The ADICO grammar allows us to distinguish between shared strategies, norms, and rules in the following way:

- *Shared strategies* have the grammar AIC. This means that they have no deontic operator and no sanction. Instead, they denote a shared strategy that agents follow in pursuing their goals. Agents follow such strategies for prudent reasons pertaining to their (perceived) self-interest. An example is: Pedestrians (A) use an umbrella to avoid getting wet (I) when it is rains (C).[4]
- *Norms* have the grammar ADIC. They thus add a deontic operator to shared strategies, but they lack sanctions if they are not followed. They are usually not followed for reasons of self-interest but are based on shared normative expectation. The agents to whom a norm applies are (normatively) expected to follow the norm, although there are no (explicit) sanctions for ignoring it. An example is: Residents (A) must (D) greet their neighbors (I) in this neighborhood (C).
- *Rules* have all the five elements: ADICO. In addition to having a deontic operator, they also contain a sanction if the institution is not followed. An example is: Car drivers (A) must (D) drive on the right side of the road (I) in the Netherlands (C), otherwise they will be fined by the police (O).
- In all three cases, institutions are based on *shared expectations*, though these are somewhat different for each case. For rules, the shared expectation is—at least partly—the expectation of a sanction if the aim (I) is not followed or achieved. For norms, the (explicit) sanction is absent. There may still be an implicit sanction in that people who do not follow the rule may be perceived as (morally) wrong or deviant. However, the main motivation for following the norm may be that people believe it to be (morally) right or (socially) appropriate, rather than the threat of

[4] Some scholars may consider this a common strategy rather than a shared strategy (Ghorbani et al. 2013). Shared strategies require a shared descriptive expectation, while common strategies do not.

sanction. Finally, for shared strategies, the expectation is not normative ("other agents are supposed to behave in a certain way") but descriptive ("other agents will most likely behave in a certain way"). Given this descriptive expectation, an agent may follow the shared strategy out of (perceived) self-interest.

We may now account for the embedding of values in institutions in a similar way to how we understood the embedding of values in technological artifacts. Substituting institution for technological artifact in the earlier account results in the following account:

> Institution R embodies value V if R is conducive to V because R has been designed for V.

The phrase "R is conducive to V" may be understood as follows:

> Institution R is conducive to value V if V is achieved (or respected) when R is followed by all relevant agents under the appropriate conditions.

Again, a few examples can illustrate this account:

1. The first example is the previously mentioned rule, "Car drivers (A) must (D) drive on the right side of the road (I) in the Netherlands (C), otherwise they will be fined by the police (O)." This institutional rule embodies the value of (traffic) safety because if all agents follow this rule, it will (under normal circumstances) be conducive to traffic safety. Moreover, this institutional rule has been deliberately designed to achieve traffic safety.
2. The previously mentioned norm, "Residents (A) must (D) greet their neighbors (I) in this neighborhood (C)," is another example. This norm may be said to embody the value of politeness because it is conducive to politeness (under normal circumstances) and such norms are typically brought into being, that is, they are "designed," either implicitly or explicitly, to serve the value of politeness.
3. For an example of a shared strategy, consider the case of people walking on the right side of the pavement to avoid bumping into those going in the opposite direction. In terms of the ADICO grammar, this strategy could be formulated as follows: "Pedestrians (A) walk on the right side of the pavement (I) in busy city centers (C)." The embodied value may here be something like convenience. Again, if everyone follows the shared strategy, that is conducive to the value of convenience. And at least one of the reasons why people adapt ("design") such shared strategies is because it serves the value of convenience.

Similar to the case in which technical artifacts embody values, an institution only embodies a value if both conditions (design and conduciveness) are met and are connected. So, an institution that has been designed for a certain value V, but which—when followed by all relevant agents—does not contribute to achieving V, does not embody V. Similarly, an institution that somehow turns out to unexpectedly contribute to a value that it was never deliberately designed for does not embody that value. Still, this value may be embedded in the institution through redesign if it

is deemed desirable, or, if a disvalue is realized, the institution may be redesigned to embed an opposite, positive, value.

Again, as in the case of technical artifacts, values embodied in an institution are not always realized. One reason might be that an insufficient number people actually follow the institution in practice, so the value is not realized. This may, for example, happen in the case of so-called empty institutions, that is, institutions that are not (or hardly) followed in practice (Ho 2016).

7 Human Agents

In so far as human actions are required for the proper functioning of sociotechnical systems, human agents can be seen as being a part of such systems, fulfilling various roles such as user, operator, and designer. As users, they will often not use the entire system but rather specific technical artifacts within the system, and in doing so, they may follow the artifact's use plan, if it aligns with their own goals. As operators, they will not just use parts of the system but will also monitor the functioning of the system, or at least relevant parts of it, and they may adjust their behavior to guarantee the proper functioning of the system. Lastly, as designers, they may seem external to the system. However, the continued existence and performance of a sociotechnical system may require continuous redesign; in that sense, then, designers may also be seen as part of the sociotechnical system.

Human agents, particularly in operator roles, will typically combine an internal perspective with an external one (Franssen 2015). So, human agents may operate from within the system, for example following existing use plans or operator instructions (i.e., existing institutions), but at the same time, they fulfill other roles in other systems or in society. In addition, they are individuals with moral agency who autonomously reflect on their roles and values. This combination of perspectives makes human users and operators a liability for system designers, as they may deviate from their prescribed roles and thus endanger the functioning of the system as well as the realization of the values the designers embedded in that system. However, the fact that human agents can take an external perspective and can reflect implies that they are able to improvise in unexpected situations and can therefore contribute to the proper functioning of the system and the achievement of certain values, which would not have been realized without their intervention.

Since sociotechnical systems contain institutions and technical artifacts that embody certain values, the behavior of human agents in the context of such systems will be different than in other contexts. Nevertheless, it is likely that the values of human agents will also influence how they behave in the context of sociotechnical systems. This is indicated schematically in Fig. 2.

A human agent acts with a certain technical artifact. This acting is based on the relevant institutions and the values embodied in them (V_I), as well as on the values embodied in the technology (V_T) and the (personal) values of the agent (V_A). This action is an intentional-causal relation. It is intentional because the agent acts based on certain intentions and values. It is causal because the agent (in most cases) does some physical activity with the artifact, which may be understood in causal terms.

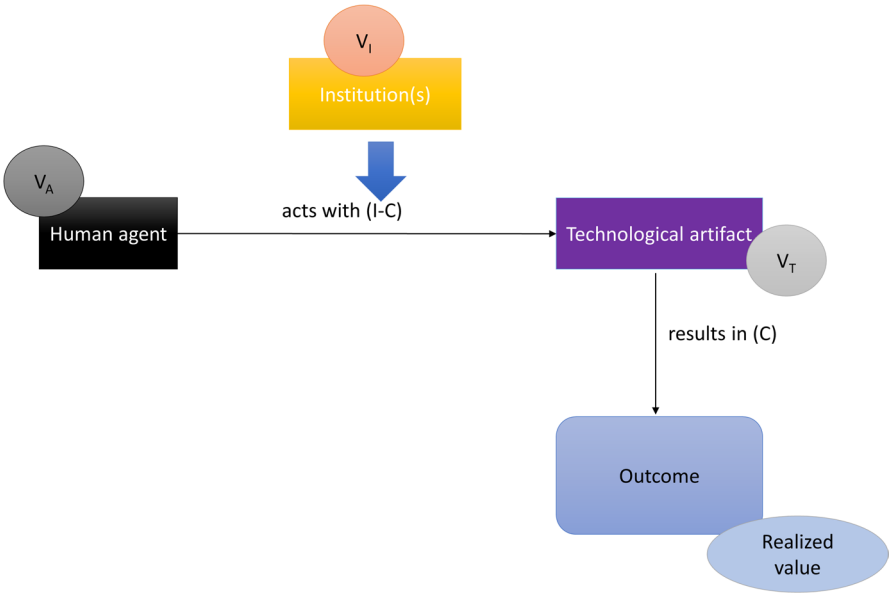

Fig. 2 Human agents acting within a sociotechnical system. V_A are the values of the agent, V_I the values embodied in the relevant institution(s), and V_T the values embodied in the technological artifact. I-C stands for intentional-causal, and C for causal

This acting then results in certain consequences, which are also based on the values embodied in the technical artifact This outcome determines the realized value, which may be different from V_A, V_T, or V_I.

Because human agents are reflective, they will monitor and evaluate the outcomes of the sociotechnical system and compare them with their own values and the values embodied in the system. As a consequence, they may try to change the system's outcomes, either by changing their own behavior or by modifying or redesigning other elements of the sociotechnical system. To what extent they will do the latter will partly depend on their specific role in the sociotechnical system. Users, for example, are usually supposed to use the (components of the) sociotechnical system without changing it. But other agent roles may allow for changing or redesigning parts of the system. And even actors without such a role may (try to) break the existing institutions and presume a role of "moral entrepreneur" (Becker 1963) who tries to change the rules in the sociotechnical system based on their own (moral) values.

Adding this dimension provides a slightly more complicated picture (see Fig. 3).

8 Artificial Agents

Some (but not all) of the roles played by human agents in sociotechnical systems may be taken over by AAs. AAs are computer and robot systems that combine autonomy with interactivity and adaptability (Floridi and Sanders 2004). This combination could make it possible, at least in principle, to design AAs that

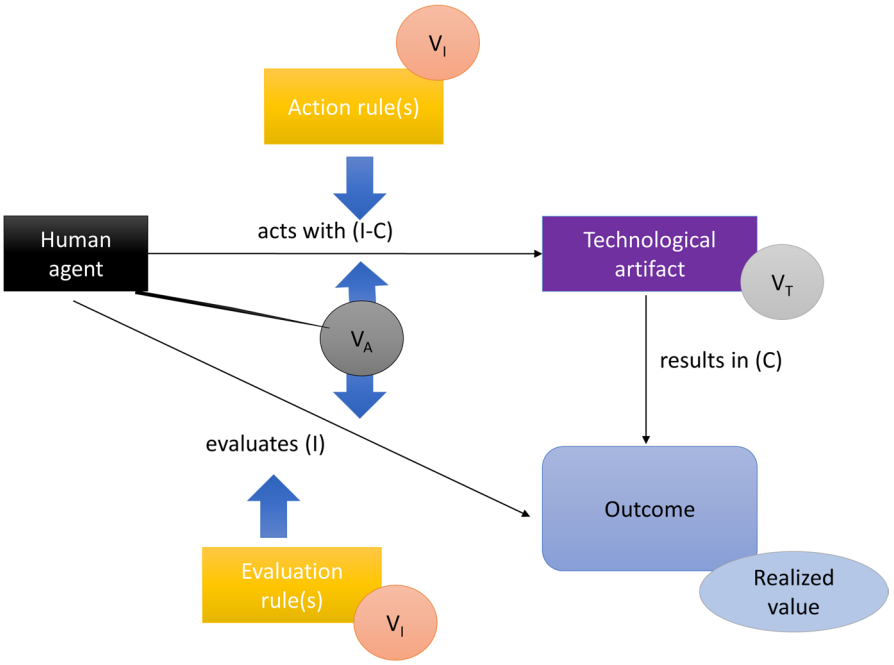

Fig. 3 Human action and evaluation in the context of sociotechnical systems. V_A are the values of the agent, V_I the values embodied in the relevant institutions, and V_T the values embodied in the technological artifact. I-C stands for intentional-causal, I for intentional, and C for causal

autonomously respect certain values in their functioning and that are able to "evaluate" certain measurable outcomes of the system and make necessary adaptations to the system: that is, the AA can adapt its own behavior or other elements of the system. Such AAs could play various roles similar to human agents, like that of operator, in a sociotechnical system.

However, there are also distinct differences between human and artificial agents. AAs are designed, whereas humans are not. Thus, artificial agents can embody values, while it would be a category mistake to say that humans embody values. (Of course, humans can have or form values, and they can instill values in other humans.) Conversely, while humans can embed values in other entities, artificial agents lack that ability as they have no intentionality. In these two respects, AAs are similar to technical artifacts, but they are also dissimilar due to their autonomy, interactivity, and adaptability. The main similarities and differences are summarized in Table 2.

Unlike technical artifacts, AAs can adapt their behavior on the basis of external inputs and interactions. This quality may strengthen as well as undermine the realization of the initially embodied values. It may strengthen it because, unlike technical artifacts, AAs can adapt to unexpected circumstances or new contexts in order to help realize their embodied values. However, the same adaptive qualities might also mean that an AA can actually "abandon" its initially embodied values and, as it were, disembody them. We say that an AA disembodies value V if it adapts itself in

Table 2 Differences between technical artifacts, artificial agents, and human agents

	Technical artifacts	Human agents	Artificial agents
Can embody values	Yes	No	Yes
Can embed values (in other entities)	No	Yes	No
Autonomous, adaptive and interactive	No	Yes	Yes

such a way that it is no longer conducive to V (under normal circumstances), even if it originally embodied V.

How an AA will use its adaptive capacities will depend in part on its specific design and on the specific way that values have, or have not, been embodied in it. Here, James Moor's (2006) taxonomy for different types of ethical agents is helpful:

1. *Ethical impact agents* are robots and computer systems that ethically impact their environment.
2. *Implicit ethical agents* are robots and programs that have been programmed (by humans) to behave according to certain values.
3. *Explicit ethical agents* are machines that can represent ethical categories and that can reason (in machine language) about them.
4. *Full ethical agents* also possess characteristics often considered crucial for human or moral agency, such as consciousness, free will, and intentionality.

Whether AAs can ever be designed as full ethical agents that possess metaphysical properties such as intentionality, free will, consciousness, moral agency, feeling, and the like, may be doubtful. Today's artificial ethical agents, at least, still seem very far removed from anything that would amount to being full ethical agents (Winfield 2019; Müller 2020). Currently, AAs may thus be designed as ethical impact agents, implicit ethical agents, or explicit ethical agents. When they are designed as ethical impact agents, they do not embody values, as their ethical impacts are not due to intentional design choices, which is one of the conditions for value embodiment. This does not mean that they are ethically irrelevant, but their ethical and value implications depend on use and idiosyncratic circumstances rather than on deliberate design choices.

Moor's implicit ethical agents embody certain values because they have been designed (by human designers) for those values, and they will respect those value when properly designed and used. However, in so far as these implicit ethical agents are autonomous, interactive, and adaptive, it is conceivable that they will develop themselves in such a way that, at some point, they are no longer conducive to the values initially embedded in them by their designers (cf. Grodzinsky et al. 2008). If human designers want to prevent this possibility, they should probably either build in certain restrictions on how such artificial agents can adapt themselves or monitor the development of such artificial agents and redesign them when necessary.

Explicit ethical agents can represent values, and other moral notions, and can "reason" about them. However, this does not necessarily mean they have values

embodied in them—that greatly depends on whether they are designed top-down or bottom-up (or in a hybrid way) (cf. Cervantes et al. 2020; Allen et al. 2005). Those designed top-down are based on a certain ethical theory or value system, and they typically have some values embodied. However, those designed bottom-up acquire their values from or in interaction with their environment. Consequently, they have (initially) no values embodied in them, and they might run the risk of picking up, or learning, not only values but also disvalues from their environment.

Explicit ethical agents seem to have two potential advantages over implicit ethical agents when it comes to embedding and realizing values in AI systems. First, because they can explicitly represent values, it seems as if it would be more straightforward to design them in such a way that they are prevented from disembodying certain values through learning and adaptation. Second, they may be better able to figure out how a value can be upheld in unexpected circumstances or new contexts, due to their "reasoning" capabilities. Of course, whether these potential advantages are realized largely depends on both the quality of the value system or ethical theory programmed into, or acquired by, the AA and the possibility of actually building ethical and context sensitivity into such an agent. Both are still huge challenges, as there is no agreement in philosophy about what the right ethical theory or value system is, and it is still not possible (if it ever will be) to provide AAs with something like ethical sensitivity (cf. Wallach and Allen 2009; Cave et al. 2019). Moreover, relying too much on explicit ethical agents for embedding values in AI systems may also have big disadvantages such as undermining human moral autonomy and responsibility (van Wynsberghe and Robbins 2019).

9 Norms in Artificial Multi-agent Systems

Whereas the behavior of and interactions between human agents are regulated by social institutions, the behavior of and interactions between AAs are regulated by (computer) code. Institutions cannot directly regulate AA behavior nor can code directly regulate human behavior. But indirect regulation is possible in both cases (see Table 3). As Lessig (1999) pointed out, code or architecture—broadly conceived—may be seen as a mode for regulating human behavior, as do social norms, laws, and the market. Of course, this does not mean that human agents execute computer code, but rather that code or architecture as it has become embodied in, for example, technological artifacts encourages or discourages certain human behaviors

Table 3 Modes of regulation of human and artificial agents in sociotechnical systems

	Institutions	Code
Human agent	Directly through social expectations	Indirectly (technical design may encourage or discourage certain human behavior)
Artificial agent	Only indirectly (institutions may be translated into technical norms)	Directly through technical norms

(see also Akrich 1992; Latour 1992; Thaler and Sunstein 2009; Fogg 2003; Norman 2000).

Conversely, social institutions may be translated into computer code that regulates the behavior of and interactions between AAs (Leenes and Lucivero 2014). This is particularly done in the (research) field of normative multi-agent systems, in which it has become common to design artificial multi-agent systems that are composed of not only AAs but also (encoded) norms (see, e.g., Hollander and Wu 2011; Singh 2014; Dybalova et al. 2014; Boissier 2006; Mahmoud et al. 2014). To distinguish such norms from social norms (and institutions), I call them *technical norms*, although they are often not technical at the semantic level.[5]

Technical norms can be created in multi-agent systems in basically two ways: through offline design or through designing AAs to autonomously discover, invent, or spread norms (Hollander and Wu 2011). In the first case, norms are specified and encoded in the agents by the (human) system designers. In the second case, agents may pick up norms from their environment in various ways, or norms may emerge from agents' mutual interactions or interactions with human agents.

In the area of multi-agent systems, various conceptualizations of norms can be found, often inspired by literature in the social sciences, law, and philosophy (see, e.g., Mahmoud et al. 2014; Aldewereld and Sichman 2013). In addition, various architectures exist that can integrate norms in multi-agent systems (Mahmoud et al. 2014). One difference between social institutions and technical norms is that while we can probably not design institutions that human agents follow without exception, it is in principle possible to design norms that artificial agents always follow. System designers may, however, prefer to design norms that allow for exceptions and that are instead used as input for "deliberations" or for weighing the advantages and disadvantages of rule following (like human agents often appear to do) (cf. Panagiotidi et al. 2013; Dybalova et al. 2014). If we want artificial agents to follow the law, we need to leave room for interpretation and situation awareness (Leenes and Lucivero 2014).

Notwithstanding differences in how technical norms are specifically implemented in AI systems, the value-embedding account can also be applied to a technical norm as follows:

> Technical norm N embodies value V if (1) N has been designed (by the human system designers) for V and (2) the execution of N within the system is conducive to V.

[5] These are to be distinguished from technical standards such as ISO standards, which are, in the terminology used here, more like institutions (but of course relevant to value embedding).

10 Values Embodied in AI Systems

So far, value-embedding has been discussed at the component level, but how can we understand it at the system level? We start with the (hypothetical) case that an AI system is designed in its entirety. Applying the general account would give the following:

> Value V is embodied in sociotechnical system S if S is conducive to V because S has been designed for V.

We may understand this as involving two conditions:

1. The system S has been designed for the value V.
2. If all the relevant institutions (including use plans for the relevant artifacts) and technical norms are followed by the human and artificial agents in the system, value V is realized.

Interestingly, this account does *not* require that for sociotechnical system S to embody value V, all the designed elements of S must also embody V. In fact, the only requirement is that following the relevant institutions (including use plans) and technical norms is conducive to V. This requires that V is embodied in some of the relevant building blocks, but it does not imply that it is embodied in all components of the system. Nor does it imply that all human agents have V as an agent (personal) value, as the values agents follow in the context of a sociotechnical system may be different from their personal values.

For many sociotechnical systems, including AI systems, condition 1 will not be met because such systems are often not designed in their entirety (e.g., Bowker et al. 2010). The difference between system components (technical artifacts, institutions, artificial agents, and technical norms) and entire sociotechnical systems is that the first are entities that are often designed in their entirety, while the second are typically not designed in their entirety but instead emerge or evolve gradually from existing systems.

Given that AI systems are often not completely designed, one might want to look for ways to relax the condition "The system S has been designed for the value V" while somehow maintaining connection with the designer's intentions. One possibility is to change this condition to "Some components of system S have been designed for value V." However, one would also want to require that system S is conducive to value V because of those components that have been designed for V, rather than because of other components or other reasons. This gives the following account:

> Value V is embodied in sociotechnical system S if S is conducive to V because of those components of S that have been designed for V.

Here, being conducive to can be understood in the same way as before, that is, V is realized when the relevant institutions and technical norms are followed.

This account allows us to say that many sociotechnical and AI systems embody certain values, especially if we apply a broad notion of (re)design. Another

interesting feature of this account is that to embed a value in an existing sociotechnical system, we need not completely redesign that system; instead, it may be enough to redesign some of its components so that it starts to embody value V. Which components would be the best candidates for such redesign would depend on the specific case and circumstances. In many cases, however, institutions would be a plausible candidate because they seem to play a particularly important role in keeping the system together. Institutions regulate how human agents interact and how they act with technical artifacts. They also play a role in how human agents evaluate system outcomes and consequently adapt their behavior (see Figs. 2 and 3).[6] This role of institutions also seems to imply an interesting lesson regarding AI systems. In AI systems, technical norms play an important role in regulating the AAs, and it might well be that if we want to embed certain values in an AI system, we should focus on those technical norms rather than (only) on the AAs themselves. This would seem to imply an important shift in focus compared to current research that often seems primarily aimed at building values and ethics into AAs, while more or less neglecting the other components of an AI system.

11 Conclusion and Lessons for Embedding Values in AI Systems

This proposed account for embedding values in AI systems may be considered a first step toward the development of further theories, approaches, and methodologies to verify, or at least assess, whether certain AI systems embody certain values, which seems crucial if we are to take the calls for respecting certain values in the design and deployment of AI and other technologies seriously. Although it is only a first step, it may be an essential step because, before we can develop methods for assessing the embodiment of values in AI systems, we must first have an account of what it means to say that a sociotechnical system embodies a value. In ending, I want to point out two tentative lessons that I think may be drawn for better embedding values in AI systems.

The first lesson relates to the difference between AI systems and more traditional sociotechnical systems. AI systems are autonomous, adaptive, and interactive, which means that they acquire many of their features during operation and due to the way they evolve rather than through their initial design. To some extent, this is also true of traditional sociotechnical systems that usually acquire (emergent) properties during their evolution that were never intended by the initial system designers. However, AI systems offer unique value-embedding opportunities and constraints because they contain additional building blocks compared to traditional sociotechnical systems. While these allow new possibilities for value embedding, they also impose constraints and risks, e.g., the risk that an AI system disembodies certain values due to how it evolves. This means that for AI systems, it is crucial to monitor their realized values and to undertake continuous redesign activities. Doing so is also crucial to deal with the unintended and unforeseen consequences of AI systems.

[6] I thank one of the anonymous reviewers for this suggestion.

While any sociotechnical system may, and will, have unintended consequences, they are probably more endemic in AI systems due to their learning capabilities. Monitoring and redesign can ensure that these systems embody the values we consider important as a society (whether those are values that were deemed important during the initial design of the system or values that have become important over time, e.g., due to unintended consequences). In part, this may be achieved within the AI system. For example, AAs can be tasked with monitoring value consequences and interfering when undesired consequences occur. However, given the current capabilities of artificial moral agents, it would seem wise to also ensure a role for human agents here. The need for human oversight may be considered somewhat of a paradox. On the one hand, AI systems are much more autonomous and adaptive than traditional sociotechnical systems. They therefore may seem to require much less human intervention. Nevertheless, ensuring that the right values remain embodied in them would, at least currently and in the foreseeable future, require continuous human oversight and redesign. It consequently also requires that AI systems are designed so that they can remain under meaningful human control (Santoni de Sio and van den Hoven 2018).

The second lesson relates to *how* we can embed values in AI systems. Here, attention needs to be given to both the system and the component levels. Much attention is currently being given to machine ethics and the possibility of designing artificial moral agents (e.g., Wallach and Allen 2009; Anderson and Anderson 2011). However, focusing solely on the possibility of embedding values in AAs without looking at the other AI system components and the effects at the system level is too limited and might even be misleading. For example, designing explicit ethical AAs may offer new possibilities for embedding values in AI systems, but it would also introduce new risks. Therefore, as I suggested in this paper, sometimes technical norms may be a better target than AAs for embedding values in AI systems because those norms regulate the behavior of the AAs in the system. One might even think of making some of the technical norms in an AI system unmalleable to decrease the risk of values getting disembodied as the system evolves.

Acknowledgements 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 program under grant agreement No 788321. I thank the members of the ValueChange project team, in particular Amineh Ghorbani, Anna Melnyk and Michael Klenk, for comments on an earlier version. I would also thank two anonymous reviewers for their comments, which have significantly improved the paper. I thank Sheri Six for English corrections. An earlier version was presented at the meeting of the Society for the Philosophy of Technology (SPT) in 20–22 May 2019 in Texas (USA).

References

Akrich, M. (1992). The Description of Technical Objects. In W. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 205–224). Cambridge: MIT Press.

Aldewereld, H., & Sichman, J. S. (2013). *Coordination, organizations, institutions, and norms in agent systems VIII: 14th International Workshop, COIN 2012 Lecture notes in artificial intelligence* (Vol. 7756). New York: Springer.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155. https://doi.org/10.1007/s10676-006-0004-4.

Anderson, E. (1993). *Value in ethics and economics*. Cambridge: Harvard University Press.

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

Bauer, J. M., & Herder, P. M. (2009). Designing Socio-Technical Systems. In A. Meijers (Ed.), *Philosophy of Technology and Engineering Sciences* (pp. 601–630). Amsterdam: Elsevier.

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers, 23*(1), 4–17. https://doi.org/10.1016/j.intcom.2010.07.003.

Becker, H. S. (1963). *Outsiders. Studies in the Sociology of Deviance*. New York: The Free Press of Glencoe.

Behymer, K. J., & Flach, J. M. (2016). From Autonomous Systems to Sociotechnical Systems: Designing Effective Collaborations. *She Ji: The Journal of Design, Economics, and Innovation, 2*(2), 105–114. https://doi.org/10.1016/j.sheji.2016.09.001.

Bicchieri, C. (2006). *The grammar of society: the nature and dynamics of social norms*. New York: Cambridge University Press.

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence research*. New York: Springer.

Boissier, O. (2006). *Coordination, organizations, institutions, and norms in multi-agent systems: AAMAS 2005 International Workshops on Agents, Norms and Institutions for Regulated Multi-Agent Systems, ANIREM 2005, and From Organizations to Organization-Oriented Programming in Multi-Agent Systems, OOOP 2005* (Lecture notes in computer science,, Vol. 3913). Berlin; New York: Springer.

Borenstein, J., Herkert, J. R., & Miller, K. W. (2019). Self-driving cars and engineering ethics: The need for a system level analysis. *Science and Engineering Ethics, 25*(2), 383–398. https://doi.org/10.1007/s11948-017-0006-0.

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2010). Toward information infrastructure studies: Ways of knowing in a networked environment. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of internet research* (pp. 97–117). Dordrecht: Springer.

Brosch, T., Sander, D., Clément, F., Deonna, J. A., Fehr, E., & Vuilleumier, P. (2016). *Handbook of value: perspectives from economics, neuroscience, philosophy, psychology and sociology*. Oxford: Oxford University Press.

Bruijn, H., & Herder, P. M. (2009). System and actor perspectives on sociotechnical systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 39*(5), 981–992. https://doi.org/10.1109/TSMCA.2009.2025452.

Calvert, R. L. (1995). The rational choice theory of social institutions: cooperation, coordination, and communication. In E. A. Hanushek & J. S. Banks (Eds.), *Modern political economy: Old topics, new directions (Political economy of institutions and decisions)* (pp. 216–268). Cambridge: Cambridge University Press.

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.

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics, 26*(2), 501–532. https://doi.org/10.1007/s11948-019-00151-x.

Coeckelbergh, M. (2020). *AI ethics*. Cambridge: The MIT Press.

Crawford, S. E. S., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review, 89*(3), 582–600. https://doi.org/10.2307/2082975.

Dam, K. H., Nikolic, I., & Lukszo, Z. (2013). *Agent-based modelling of socio-technical systems* (Vol. 9). New York: Springer.

Dennett, D. C. (1987). *The intentional stance*. Cambridge: MIT Press.

Dybalova, D., Testerink, B., Dastani, M., & Logan, B. A Framework for Programming Norm-Aware Multi-agent Systems. In *Cham, 2014* (pp. 364–380, Coordination, Organizations, Institutions, and Norms in Agent Systems IX): Springer International Publishing

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying values in technology. Theory and practise. In J. Van den Hoven & J. Weckert (Eds.), *Information technology and moral philosophy* (pp. 322–353). Cambridge: Cambridge University Press.

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

Fogg, B. J. (2003). *Persuasive technology: using computers to change what we think and do (The Morgan Kaufmann series in interactive technologies)*. Amsterdam: Morgan Kaufmann Publishers.

Franssen, M. (2015). Design for values and operator roles in sociotechnical systems. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of ethics, values, and technological design: sources, theory, values and application domains* (pp. 117–149). Dordrecht: Springer.

Franssen, M. Modelling Systems in Technology as Instrumental Systems. In *Berlin, Heidelberg, 2014* (pp. 543–562, Model-Based Reasoning in Science and Technology): Springer Berlin Heidelberg

Geels, F. W. (2004). From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research Policy, 33*(6), 897–920. https://doi.org/10.1016/j.respol.2004.01.015.

Ghorbani, A., Aldewereld, H., Dignum, V., & Noriega, P. Shared Strategies in Artificial Agent Societies. In *Berlin, Heidelberg, 2013* (pp. 71–86, Coordination, Organizations, Institutions, and Norms in Agent Systems VIII): Springer Berlin Heidelberg

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 10*(2), 115–121. https://doi.org/10.1007/s10676-008-9163-9.

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

Hirose, I., & Olson, J. (2015). *The Oxford handbook of value theory*. New York: Oxford University Press.

Ho, P. (2016). Empty institutions, non-credibility and pastoralism: China's grazing ban, mining and ethnicity. *The Journal of Peasant Studies, 43*(6), 1145–1176. https://doi.org/10.1080/03066150.2016.1239617.

Hollander, C. D., & Wu, A. S. (2011). The current state of normative agent-based systems. *Journal of Artificial Societies and Social Simulation, 14*(2), 6. https://doi.org/10.18564/jasss.1750.

Houkes, W., & Vermaas, P. E. (2010). *Technical functions: on the use and design of artifactartifacts (Philosophy of engineering and technology* (Vol. 1). Dordrecht: Springer.

Houkes, W., Vermaas, P. E., Dorst, K., & de Vries, M. J. (2002). Design and use as plans. An action-theoretical account. *Design Studies, 23,* 303–320.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. (1st ed.): IEEE.

Ihde, D. (2012). *Experimental phenomenology: multistabilities* (2nd ed.). Albany: State University of New York Press.

Illies, C., & Meijers, A. (2009). Artifact artifacts without agency. *The Monist, 92*(3), 420–440.

Jacobson, D. (2011). Fitting Attitude Theories of Value. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8,* 195–205.

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2), 123–133. https://doi.org/10.1007/s10676-008-9174-6.

Jones, A. J. I., Artikis, A., & Pitt, J. (2013). The design of intelligent socio-technical systems. *Artificial Intelligence Review, 39*(1), 5–20. https://doi.org/10.1007/s10462-012-9387-2.

Klenk, M. (2020). How do technological artifact artifacts embody moral values? *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00401-y.

Kroes, P. (2010). Engineering and the dual nature of technical artifactartifacts. *Cambridge Journal of Economics, 34*(1), 51–62. https://doi.org/10.1093/cje/bep019.

Kroes, P., Franssen, M., Van de Poel, I., & Ottens, M. (2006). Treating socio-technical systems as engineering systems: some conceptual problems. *Systems Research and Behavioral Science, 23*(6), 803–814. https://doi.org/10.1002/sres.703.

Kroes, P., & Meijers, A. (2006). The dual nature of technical artifactartifacts. *Studies In History and Philosophy of Science Part A, 37*(1), 1–4.

Kroes, P., & Verbeek, P.-P. (Eds.). (2014). *The moral status of technical artifactartifacts*. Dordrecht: Springer.
Latour, B. (1992). Where are the missing masses? In W. Bijker & J. Law (Eds.), *Shaping Technology/ Building Society; Studies in Sociotechnical change* (pp. 225–258). Cambridge: MIT Press.
Latour, B. (1993). *We have never been modern*. New York: Harvester Wheatsheaf.
Leenes, R., & Lucivero, F. (2014). Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design. *Law, Innovation and Technology, 6*(2), 193–220. https://doi.org/10.5235/17579961.6.2.193.
Lessig, L. (1999). *Code and other laws of cyberspace*. New York: Basic Books.
Mahmoud, M. A., Ahmad, M. S., Mohd Yusoff, M. Z., & Mustapha, A. (2014). A review of norms and normative multiagent systems. *The Scientific World Journal, 2014,* 684587. https://doi.org/10.1155/2014/684587.
Miller, B. (2020). Is Technology Value-Neutral? *Science, Technology and Human Values*. https://doi.org/10.1177/0162243919900965.
Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21. https://doi.org/10.1109/MIS.2006.80.
Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2020 Edition)*.
Nickel, P. J. (2013). Trust in Technological Systems. In M. J. de Vries, S. O. Hansson, & A. W. M. Meijers (Eds.), *Norms in technology* (pp. 223–237). Dordrecht: Springer.
Norman, D. A. (2000). *The design of everyday things*. Cambridg: MIT Press.
North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge: Cambridge University Press.
Ostrom, E. (2005). *Understanding institutional diversity (Princeton paperbacks)*. Princeton: Princeton University Press.
Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, games, and common-pool resources*. Ann Arbor: University of Michigan Press.
Ottens, M., Franssen, M., Kroes, P., & Van de Poel, I. (2006). Modeling engineering systems as sociotechnical systems. *International Journal of Critical Infrastructures, 2,* 133–145.
Panagiotidi, S., Vázquez-Salceda, J., & Dignum, F. Reasoning over Norm Compliance via Planning. In *Berlin, Heidelberg, 2013* (pp. 35–52, Coordination, Organizations, Institutions, and Norms in Agent Systems VIII): Springer Berlin Heidelberg
Pasmore, W. A., & Sherwood, J. J. (1978). *Sociotechnical systems : a sourcebook*. La Jolla: University Associates.
Peterson, M., & Spahn, A. (2011). Can technological artifactArtifacts be moral agents? *Science and Engineering Ethics, 17*(3), 411–424.
Pitt, J. C. (2014). Guns don't kill, people kill"; values in and/or around technologies. In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artifactartifacts* (pp. 89–101). Dordrecht: Springer.
Raz, J. (1999). *Engaging reason. On the theory of value and action*. Oxford: Oxford University Press.
SantonideSio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*. https://doi.org/10.3389/frobt.2018.00015.
Scanlon, T. M. (1998). *What we owe to each other*. Cambridge: Harvard University Press.
Searle, J. R. (1984). *Minds, brains, and science (The 1984 Reith lectures)*. Cambridge: Harvard University Press.
Singh, M. P. (2014). Norms as a basis for governing sociotechnical systems. *ACM Trans. Intell. Syst. Technol., 5*(1), 1–23. https://doi.org/10.1145/2542182.2542203.
Stevenson, C. L. (1944). *Ethics and language*. New Haven: Yale University Press.
Sullins, J. P. (2006). When is a robot a moral agent. *International Review of Information Ethics, 6*(12), 23–30.
Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: improving decisions about health, wealth, and happiness (Rev. and expanded ed.*. New York: Penguin Books.
Ullmann-Margalit, E. (1977). *The emergence of norms (Clarendon library of logic and philosophy)*. Oxford Eng: Clarendon Press.
Van de Poel, I., & Kroes, P. (2014). Can technology embody values? In P. Kroes & P.-P. Verbeek (Eds.), *The moral status of technical artifacts* (pp. 103–124). Dordrecht: Springer.
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.* Paper presented at the Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA,

Verbeek, P.-P. (2011). *Moralizing technology: understanding and designing the morality of things*. Chicago, London: The University of Chicago Press.

Vermaas, P. E., & Houkes, W. (2006). Technical functions: a drawbridge between the intentional and structural natures of technical artifacts. *Studies In History and Philosophy of Science Part A, 37*(1), 5–18.

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

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

Winner, L. (1977). *Autonomous Technology. Technics-out-of-Control as a Theme in Political Thought*. Cambridge: MIT Press.

Winner, L. (1980). Do artifacts have politics? *Daedalus, 109,* 121–136.

Zimmerman, M. J. (2015). Value and Normativity. In I. Hirose & J. Olson (Eds.), *The Oxford handbook of value theory*. Oxford: Oxford University Press.