AI & SOCIETY
https://doi.org/10.1007/s00146-023-01852-5

MAIN PAPER

The sociotechnical entanglement of AI and values

Deborah G. Johnson[1] · Mario Verdicchio[2,3]

Received: 9 October 2022 / Accepted: 14 December 2023

Abstract

Scholarship on embedding values in AI is growing. In what follows, we distinguish two concepts of AI and argue that neither is amenable to values being 'embedded'. If we think of AI as computational artifacts, then values and AI cannot be added together because they are ontologically distinct. If we think of AI as sociotechnical systems, then components of values and AI are in the same ontologic category—they are both social. However, even here thinking about the relationship as one of 'embedding' is a mischaracterization. The relationship between values and AI is best understood as a dimension of the relationship between technology and society, a relationship that can be theorized in multiple ways. The literature in this area is consistent in showing that technology and society are co-productive. Within the co-production framework, the relationship between values and AI is shown to be generative of new meaning. This stands in stark contrast to the framework of 'embedding' values which frames values as fixed things that can be inserted into technological artifacts.

Keywords Artificial Intelligence · Co-production · Design · Sociotechnical systems · Technology · Values

1 Introduction

Scholarship on values and AI is growing (Greene et al. 2016; Torresen 2018; Chatila and Havens 2019; Umbrello 2019; Coeckelbergh 2020; Kop 2020; van de Poel 2020; Salo-Pöntinen 2021). A quick survey of this scholarship suggests that the relationship between values and AI is thought of in a wide variety of different ways. This is reflected in the array of terms used to describe this relationship. Some use the language of *aligning* AI to human, social or ethical values (Shahriari and Shahriari 2017; Kim et al. 2021); others refer to *integrating* values into technical design (Umbrello et al. 2021); some write about designing *for* values; others seek to *embed* values in AI (van de Poel 2020), or *codify* ethical behavior inside machines (Arkin 2009); some even refer to designing AI that will *obey* laws and values (Etzioni & Etzioni 2016). Each of these approaches suggests that AI and values can be brought together in some sort of additive way, i.e., we can put values into AI or accurately identify them in AI code.

However, these approaches raise more questions than answers. How can a value, an abstract, social concept, be 'in' a concrete, computational entity? How can something computational—a series of 0 s and 1 s—immutably carry a human, social notion? In what follows, we put forward an account of the relationship between values and AI. Our account rejects the idea that values can be added to or embedded in AI.

To understand the relationship between values and AI, both concepts must be clarified. For clarification on values, we draw on psychology literature which suggests that values are those things that people consider important. For clarification on AI, we distinguish two concepts of AI: computational artifacts and sociotechnical systems. When AI is understood as the former, values and AI cannot be added together because they are ontologically distinct. If AI is understood as the latter, then components of AI are in the same ontologic category as values—they are both social—and we can investigate further how they relate within this category.

✉ Mario Verdicchio
mario.verdicchio@unibg.it

Deborah G. Johnson
dgj7p@virginia.edu

[1] Engineering and Society, University of Virginia, Charlottesville, VA, USA

[2] Department of Foreign Languages, Literatures and Cultures, University of Bergamo, Bergamo, Italy

[3] Berlin Ethics Lab, Technische Universität Berlin, Berlin, Germany

Published online: 27 January 2024

The relationship between values and AI can be framed as a special case of the relationship between values and technology. Hence, the scholarship in this broader area can inform our understanding of AI and values. The value sensitive design (VSD) literature is especially interesting here because it seems to presume that one can intentionally design for values. The literature on philosophy of technology is also relevant as it seeks to understand techno-moral change in a co-production framework.

The co-production thesis captures significant aspects of the relationship between values and AI. Insofar as people attribute values to things, values derive from the complicated processes by which meaning is given to things. In those processes, people come to accept particular descriptions from an almost infinite number of possible descriptions of an AI system. In this respect, values remain in human beings and are not embedded or embodied in AI or any other technical artifact.

In what follows we are interested in critiquing the idea that values can be embedded in AI especially when AI is understood as computational artifacts. Throughout our analysis we use three different examples of AI systems that have been associated with specific values: an AI system used to distribute job advertisements in an equitable way; Tay, the AI Twitter-bot that wrote tweets that were considered racist; and the Iron Dome, an AI-based anti-missile system designed to realize the value of defending Israeli lives, by detecting and taking down missiles launched by Hamas.

2 Values

In the sentence 'Joan values family and friendship', 'values' is used as a verb to express the idea that Joan considers family and friendship important. In the sentence 'The value of social justice should influence the decisions of engineers', the phrase 'the value of' qualifies social justice as something of importance, something that should be promoted or protected by engineers. A value is that which someone or some group regards as having importance or worth.

This simple account of value is found, among other places, in the field of psychology. In their chapter, *Values and the Human Being*, Cieciuch and Schwartz (2018) explain that "The semantic core of the concept of values is importance…". They cite Rokeach (1973, p. 5) as defining values as "enduring beliefs that a specific mode of conduct or end state of existence is personally or socially preferable to an opposite or converse mode of conduct or end state of existence." Schwartz (1994, p. 21) describes values as "desirable transsituational goals, varying in importance, that serve as guiding principles in the life of a person or other social entity."

The psychology literature often distinguishes between individual and social values. Individual values have been taken up by a number of scholars that seek to help individuals identify their own values. See, for example, the Giving Voice to Value movement, an action-oriented approach to values-driven leadership with the purpose of preparing people to effectively act on their values (Gentile 2010). Our interest here is in *social* values. People learn social values through social interaction within a family, culture, country, community, ethnic group, etc. Some sociological theories suggest that individual values are the result of an internalization of social values (Scott 1971).

Importantly, social values are abstract notions that need to be interpreted (Braithwaite and Blamey 1998). In this respect they are fluid. Take, for example, equality which has been interpreted in very different ways at different times and in different contexts or take universal suffrage, a valued concept that is often contested and sometimes expanded. The value of respect for persons did not have any connection to medicine until the 1970s, when its interpretation was expanded to include the necessity for doctors to obtain informed consent from their patients before using them in experiments. Privacy is also a contested concept: on the one hand, there are companies that take privacy to mean having a complex set of policies that their customers must agree to (often by default, without reading them), and on the other, there are concerned citizens and activists who consider this understanding of privacy to be a sham. Of course, sometimes there is consensus about the meaning of a specific value, e.g., the value of human life, as when an interpretation is put into law, e.g., the value of life is interpreted to mean that laws against murder are called for. Other times, discussions and contestations about a value may continue over centuries.

So, social values are not fixed or static concepts. They are fluid in interpretation, and interpretations change and vary over time. In what follows we try to show that this makes the relationship between values and AI much more complicated than is generally understood.

3 Two concepts of AI

The literature on values and AI often seems ambiguous about what AI is. On the one hand, AI is taken to be computational artifacts, and, on the other hand, it is recognized as having social components, including social values.

3.1 AI as computational artifacts

Born as a subfield of Computer Science, AI was originally conceived in terms of computational artifacts, that is, artifacts whose operation is based on computation, such as algorithms[1], programs running on computers, and programs controlling robots. Starting from algorithms, computational artifacts constitute a range of more and more complex devices in which algorithms are written and stored inside memories of computers in the form of programs commanding the input of operands, the execution of operations, and the output of the results of those operations. Computers are typically equipped with peripherals for the exchange of digital data with other computers and human users and, in the case of robotic artifacts, they have mechanical peripherals that allow them to move in and act upon the external physical environment.

Computer Science in general is about computational artifacts, and understanding AI as computational artifacts means recognizing AI as a special kind of those artifacts, namely those meant to achieve the results typically associated with human cognition, including perception, reasoning, and making decisions. This line between AI artifacts and other Computer Science artifacts is unclear and blurry because what is considered typical of human cognition is an evolving concept. Nevertheless, the understanding of AI as a computational entity is clear; in this view, AI is algorithms, programs, computers, and robots.

To illustrate how AI is understood as a computational artifact, consider the case of an AI algorithm for the distribution of advertisements. In their work, Lambrecht and Tucker (2019) analyze how job advertisements are distributed in a social network. The algorithm is designed to automate the optimization of a number of parameters characterizing advertisement delivery. When a user loads a page of the social network's website, the advertisement algorithm conducts an auction to determine which advertiser's ad will be shown to the user. The result of the auction is typically determined by the highest bid placed by an advertiser (e.g., the advertiser has previously agreed to pay at most 0.60 Euros to show their ad to a user) relative to all other bids. The algorithmic outcome of the auction also takes into account a "quality score", that is, a predictive measure of the likelihood that the user will click on that ad. In their experiment, Lambrecht and Tucker used an ad for jobs in the Science, Technology, Engineering and Math (STEM) fields that was explicitly designed to be gender-neutral. The empirical results of their analysis showed that fewer women were shown the ad than men across the social network. According to the authors, users who are younger women are a prized demographic from the perspective of product sales, and, consequently, the ad bidding war around them is more competitive resulting in ads shown to women being more expensive.

When the ad distribution system is viewed simply as a computational artifact, the resulting distribution of ads seems to come from the algorithm. From this perspective, one might think that the apparent gender-bias of the system derives from the code. This is the idea engendered by talk of "embedding values."

However, the ad distribution case reveals much more. Among other things, the AI involved an auction. As well, social ideas about gender and a desire for more women in STEM, the financial concerns of advertisers, and much more, produced the outcome. When AI designers frame AI as computational artifacts, they may only see code as the source of a value-laden outcome, and, hence, only see better code as the solution.

3.2 AI as sociotechnical systems

The elements—the social factors and actors—missed when AI is viewed purely as computational artifacts are captured when AI is understood as sociotechnical systems. STS (Science, Technology and Society) scholars argue that technologies, including AI, are more accurately and productively understood as systems in which artifactual behavior is combined with human, social, and organizational behavior (Baxter and Sommerville 2011) to produce results. Viewing AI as sociotechnical systems provides a broader scope to our understanding of AI. It does not deny or ignore the fundamental and defining contribution of computation to AI, but it takes into consideration the important relations that hold between AI artifacts and the people who design them, those who develop the data processed by the algorithm, those who make decisions on the basis of the output, and so on. Neglecting these relationships is an oversight that has been named "sociotechnical blindness" (Johnson and Verdicchio 2017).

STS scholars argue that technology and society are intertwined in profoundly deep ways. The entanglement can be understood from a number of different perspectives, and we mention only a few here that are especially relevant to AI. First, computational artifacts are designed by humans for social ends and purposes, e.g., algorithms are designed to achieve better decisions, quicker, and often on a larger scale. Better, quicker, and larger scale are all human ends, understood in particular ways by particular actors, i.e., the goals of insurance companies, government agencies, marketers, etc. Second, while computational artifacts change the configurations in machines and produce output, none of that happens without human behavior, that is, humans have to design the programs, input the programs and data into computers, and receive output. None of these actions and

[1] Whether algorithms lack the concreteness usually characterizing artifacts is an interesting discussion that lies outside the scope of this paper. Let it suffice here to point out that algorithms need at least to be given a concrete form in terms of a description in a language (natural or symbolic) for people to illustrate them, code them, modify them and use them.

their consequences, i.e., the programs, the data, the output, have significance or meaning without humans attaching meaning to them.

Framing AI as sociotechnical systems allows us to see that AI isn't just lines of code; what AI achieves is a result of both human and machine (computational) behavior. We can illustrate the difference between viewing AI as computational artifacts and viewing AI as sociotechnical systems by returning to the advertisement distribution system mentioned earlier. In this case, the algorithm, i.e., the lines of code of the algorithm, worked together with a variety of other social and non-social components of the system to produce the outcome. These include: the financial goals of the advertisers, i.e., to keep the costs of their ad campaigns low; the cultural and political context, i.e., the socio-cultural stereotypes around consumerism and gender; and the subsequent willingness of advertisers to pay higher prices for ads to be shown to women. Moreover, while the algorithm and the relevant code constitute the computational part of the AI, the purposes that led to that coding (e.g., the instructions aiming at lowering the ad cost) were generated and given meaning by people, that is, the coding itself is the result of strategic decisions by the board of the social network platform in trying to accommodate its customers. We are much more likely to be aware of and acknowledge these influences on the outcome of the AI when we frame it as a sociotechnical system.

4 The entanglement of values and AI

The relationship between values and AI can best be accounted for in the sociotechnical systems perspective. Indeed, a case can be made for claiming that values cannot be 'in' computational artifacts, and we will discuss that claim later.

When AI is viewed as a sociotechnical system, the system can be described, conceptualized, and analyzed in a multiplicity of ways using social concepts. As mentioned before, the concept of sociotechnical systems comes from the field of Science and Technology Studies (STS). The field consists of a rich literature arguing for an expanded notion of technology and against viewing technology as simply material objects or artifacts. Different authors propose particular theories for conceptualizing sociotechnical systems and provide a wide range of ways to think about the intertwining of the social and the technical in any particular system. The broadest accounts treat technology as a complex network of human and non-human actors (Latour 2005). This way of thinking about technology is open to any number of strategies to characterize the complex network. It is nearly impossible to analyze and theorize the whole, so when a sociotechnical system is conceptualized, analysts generally pick out certain aspects of the system on which to focus. The emphasis may be on users (Oudshoorn and Pinch 2003; Garrety and Badham 2004; Roy et al. 2009), ideas (Hong and Sullivan 2009; Hofstetter et al. 2021), power relations (Walker 2000; Paz et al. 2021), historical change (Misa 1988; Little 2000), regulatory regimes (Whitford and Tucker 2009; Finch et al. 2017), funding (Branscomb and Auerswald 2002; Li 2020) or values (Le Dantec et al. 2009; Lanzeni and Pink 2021). Any of these dimensions of a sociotechnical system may be highlighted in a particular analysis. Our focus here is, of course, on values.

4.1 Confusion about AI

To begin to understand the advantages of the sociotechnical system perspective for understanding values and AI, it is useful to highlight the confusion that results from *not* recognizing the distinction between the two concepts of AI. Consider, for a start, VSD. Conceived in the 1990s, VSD is described as a *theory and method* for designing technologies in ways that consider the human values that might be impacted by a design. According to Friedman and Hendry (2019), VSD positions designers and others "to make insightful investigations into technological innovations in ways that foreground the well-being of human beings and the natural world." The fundamental claim of VSD is that explicit attention to values during the design process can make a difference in how a newly designed and adopted technology affects values. Despite having grown significantly since its beginnings (van den Hoven 2013), the approach has not gone forward without criticism. Some have challenged the way VSD conceptualizes values. Borning and Muller (2012), for example, point to the tendency in VSD to think of values as universal and the failure to see them as contextualized. This criticism is consistent with our analysis of values as not fixed entities but rather social, fluid, and in need of interpretation.

Moreover, although VSD scholars are primarily interested in how to structure design *methodologies and settings* (to ensure that values are adequately considered), some scholars seem to conflate *process and product*. Arguably, this is what has happened when AI scholars talk about inserting or embedding values into AI. They seem to leap from the idea of a design process taking values into account to a product that has values 'in' it.

Part of the problem here is ambiguity about whether design refers to the design of artifacts or the design of sociotechnical systems of which artifacts are a part. Of course, one cannot design one without the other. Even if one is designing a three-dimensional object, one has in mind a context in which it will be used. In much of the literature on embedding values into AI, it is unclear whether the AI being referred to is the computational artifact or the sociotechnical system though the words themselves, i.e., inserting, embedding, embodied, tend to suggest materiality and, hence, the artifact.

An example of this confusion is found in van de Poel (2020). Van de Poel proposes a version of VSD for AI by imagining that designers go through multiple feedback loops of design and re-design to embed values in AI. Van de Poel seems to have in mind that by changing the code of an AI through feedback loops, values are put into the AI. Yet it is hard to tell whether he thinks the values are in the code or whether he recognizes that it is not the code alone but the way the code works with other social components of the system and the perception of the effects of the operation of a system. He draws a distinction between "embedded" values, which he seems to think are in the code, and "realized" values when the code is "properly used". Van de Poel warns that the two kinds of values may be very different from each other, when a technology is used in a way that was not intended or foreseen by the designers. So he seems to recognize a distinction between the intrinsic characteristics of the code and how the code relates to its surroundings once it is executed, but he does not account for this relationality in his interpretation of values.

While clarity with respect to the two concepts of AI would be helpful here, only the sociotechnical system framework seems to allow understanding of the relationship between AI and values. That is, while one can wonder whether and how values could be in strings of 0's and 1's, ambiguity disappears when we recognize that the effects of an AI system are not just the result of the operation of code (computational artifacts). Rather the effects are the result of a combination of the operation of code in machines together with a set of social relations and meanings that people attach to outcomes in a particular context. Returning again to the ad distribution system, in order to explain how the system could be thought of as gender-biased, we had to explain that advertisers highly valued ads reaching women and how that preference played out in an auction. The algorithm worked in such a way that buyers' preferences impacted results. Gender and gender bias were not in the code. They were attributed to the system because of the effects produced by the AI system and the effects were the result of the code in combination with human behavior and the complex meanings attached to those effects.

In fact, gender is not embedded in the code. Rather, some code is taken as a representation of gender. In an oversimplified binary case, the value "0" may represent "male" and the value "1" may represent "female". For such an encoding (i.e., a correspondence between items and numbers) to work, there needs to be a social consensus about the meaning of those numerical values. This alone is highly problematic because some programmers may want to include values for the representation of non-binary users, who recognize themselves neither as "male" nor as "female". Having a computational model of (gender) bias is even more difficult, since such concept is tied to the effects that the code has on the users involved once it has been executed. In other words, code alone cannot stand for any bias, since bias can only be observed when the computational system runs and gives results that need to be interpreted by the affected people. Thus, the inclusion of concepts like gender or bias in an AI artifact is only possible from a sociotechnical perspective.

4.2 Co-production

Another advantage of the sociotechnical system concept of AI is that it accommodates the fluid nature of values discussed earlier. In fact, it is not just compatible with the fluidity of values but with the co-production thesis as it applies to AI. Co-production has been proposed as a framework to counter the oft-criticized dualism that juxtaposes science and technology on the one side, and values and social arrangements on the other (Jasanoff 2004). Although the thesis applies broadly to social phenomena, Swierstra et al. (2009) have coined the term 'techno-moral change' to capture the ways that moral values influence and are influenced by technology. Drawing on Swierstra, Nickel et al. (2022) focus on the disruptive character of techno-moral change and in particular the uncertainty that is created when new technologies are adopted, and people do not know what norms and values apply.

A good example illustrating co-production in AI is the case of the now infamous Twitter-bot Tay developed by Microsoft. Tay was meant to interact with human users via Twitter and learn common habits of speech from their tweets. Due to a coordinated effort by some Internet trolls, after less than one day Microsoft had to shut the experiment down because the bot was publishing highly inappropriate tweets that included racist, sexist, and antisemitic language. Wolf et al. (2017) contends that this and other similar incidents are symptoms of a deeper problem characterizing AI software that learns, that is, programs that alter their own lines of code as a result of their interactions with other programs or human users. They argue that in the context of software that interacts with and learns from people directly or indirectly on social media platforms, the software developers have additional responsibilities to go beyond what they do in the case of standard non-learning programs and take into account all the possible results of the learning. Insofar as Wolf's analysis is focused on software and software only, the authors seem to be thinking of AI as computational artifacts and seem to forget about the contributions of users to the production of racist, sexist, and antisemitic language. It might be appropriate to say that Tay was a racist AI, but this was not because racism was embedded in the lines of code that comprise the algorithm at the heart of Tay. Rather it was because Tay—understood as a sociotechnical system—includes the humans that interacted with the algorithm and from whom Tay learned.

Moreover, the Tay example shows that co-production goes beyond just the designers and users of AI systems; it involves the wide range of social factors and actors that affect and are affected by an AI system. In Tay's case, not only co-production refers both to the social phenomena that led to the classification of Tay's tweets as "racist" (racism in society shaping Tay) but also Tay expanding the meaning of racism by providing a new technological instance of the concept. Tay led to a new meaning of the term "racism" in the form of an AI.

In this example, we can see the limitations of thinking of AI only as computational artifacts by engaging in the thought experiment of imagining how the programmers at Microsoft might modify Tay's code so as to eliminate "racism" from all its future outputs, or anything that is deemed "unacceptable" by society's standards (whichever society the programmers intend to refer to). Such results, if achievable at all, could not be achieved simply by altering the code: the task would require an extensive and up-to-date knowledge of attitudes, principles and mores in a particular society, a control over who Tay is interacting with, surveillance and censorship, and so on.

4.3 Attributing values

In both of the examples that we have discussed so far people attributed values to an AI. Some thought the ad distribution system was biased; some thought Tay was racist or exhibited a (new) form of racism. It might seem, then, that the values are in the perceptions of people and not in the AI. However, it would be a mistake to think this is an either/or matter. People have reasons for attributing values to artifacts and the artifacts have characteristics that lead them to do so. The processes by which people attribute values to artifacts are complex and myriad. A value may be attributed to a technology because of media presentations of it; because of the material properties of the technology; because of the social relations required to produce the technology; because of the impacts of use of the technology; and so on.

Recent work in philosophy of technology has focused in particular on the relationship between a technological artifact and the ways in which it may be used. One of the first conceptualizations was Ihde's notion of "multistability". According to Ihde, there are multiple intrinsic possibilities in a technological artifact, which makes it difficult if not impossible to predict all the possible consequences of the deployment of a particular artifact (Ihde 1999). While Ihde focused on "intrinsic possibilities" of technological artifacts, more recent work in the field has expanded discourse on technology to include the question of how values are 'embedded' and 'embodied" in technology. For example, reacting against the idea that technology is neutral, van de Poel and Kroes (2014) put forward an account of how values are embodied in artifacts. They do this by emphasizing the role of the designer's intentions and then make the distinction, mentioned earlier, between intended and realized values. The thrust of this is to show how values are embodied in artifacts. This focus on the properties of the artifact has continued with increasing recognition of the role of context (environment) of use in constituting values. For example, Klenk (2021) distinguishes intrinsic and extrinsic characteristics of artifacts as response-dependent. He seems to recognize that an artifact's characteristics are relational and uses the concept of affordances. According to Klenk, artifacts have intrinsic characteristics that are physical properties that may work as affordances. For example, a knife may be shaped in a certain way (i.e., one end has the shape of a handle) and this intrinsic characteristic creates an affordance in a certain environment. For example, in a kitchen where a person needs to chop some ingredients, the knife's shape makes the artifact easy to grab and use in the way it was intended. The affordance, a property that acquires meaning in that specific environment, is an extrinsic characteristic of the artifact. It is the affordance, according to Klenk, that gives the artifact value.

Klenk seems to misstep when he claims that affordances are embedded in the artifact. Affordances are relational. In the kitchen context, the knife's shape and sharpness have "affordance" only in relation to a person who grabs the knife and uses it. Affordances as such derive from a combination of the properties of artifact and the user. This misstep in Klenk's thinking seems to be recognized in Tollon (2022), who proposes a subject's specific concerns as cause for the perception of certain affordances, thus, shifting the focus away from the artifact alone and towards factors that are outside the artifact, i.e., factors that involve the people using it.

Like Tollon, De Boer (2021) recognizes the relational aspect of affordances and does not insist that they are in the artifact. De Boer focuses on the environment, understood as a "rich landscape of affordances" on which human beings allegedly "strive to have an optimal grip". De Boer claims that the form of life within which technologies are deployed influences the perception of their affordances (e.g., given the environments in which they have developed their professions, a fencing champion and a university professor will perceive different potential usages in a table). The stress on the role of the environment is aimed at developing a "situated" way of understanding how technological artifacts constitute affordances and usages.

The distinction between intrinsic and extrinsic properties of an artifact, between its materiality and its potential relations with people (forms of life as de Boer would say), may be seen as a general specification of the two concepts of AI that we distinguish, i.e., between AI as computational artifacts and AI as sociotechnical systems. The code that commands computers and robots inside their memory devices is indeed a physical entity. On the other hand, the effects of the code's execution take place in the more complex and heterogeneous network of social relations between those artifacts and people. It is in that network, and not in the physicality of the code, that values are attributed to the artifacts. In this respect, we are arguing that the relationship between values and AI can best be understood by focusing on what Klenk would refer to as the extrinsic properties of AI, but these are not embedded in a computational artifact; they are in what is understood as a sociotechnical system.

It should be noted that the focus on affordances comes with the risk of not seeing 'the forest through the trees', that is, a focus on users leads to the neglect of the many other people who can be affected by the deployment of an artifact and may attribute values to it. For example, in the ad distribution system, those who purchase the ads to recruit employees are the users. Yet the system has effects on many others including those who see the ad and those who donot see the ad, as well as the STEM workforce more broadly and social critics who see the system as biased. All of these and more may attribute values to the system.

A complete account of how and why people attribute values to AI will not be given here. Rather we will explore a case in which people attribute a variety of conflicting values to the same AI system precisely because they happen to have different relations to the system. This example also illustrates the point we made earlier about how the blurring of the two concepts of AI can lead to confusion about the relationship between values and AI.

Consider the case of the "Iron Dome," an AI anti-missile system. The recent fight between Israel and Hamas in Gaza involved airstrikes and rocket launches, and Israel's tactics for missile defense included active interception by the Iron Dome (van der Merwe 2021). The Iron Dome could be thought of as an AI system that acts within a set of pre-programmed parameters to intercept missiles. It operates using these parameters to detect a missile among other objects and to establish whether it is aimed at a high-value target. New data are gathered by the system during an engagement, and they are matched against older training data to draw new conclusions. However, the Iron Dome's performance can take a critical hit when the system is pushed to its limits by too many missiles within the system's parameters or by rockets fired outside those parameters, e.g., when missiles have too short flight times to be intercepted. Short flight times, together with very large single batteries of missiles, are tactics employed by militants to identify and exploit the limitations of the AI.

The Iron Dome can be described in a wide variety of ways that attribute a wide variety of values to it. That is, different groups of people see the Iron Dome as instantiating different, and possibly clashing, values. Some of these values may even be in conflict with the ideas of defense and security. Some may see the Iron Dome as a tool of hegemony, Western contrivance, anti-peace, etc. Generally, many different values can be associated with any particular artifact because people can see many different dimensions of technologies, depending on how they are conceptualized. Even though Israelis may see in the Iron Dome the value of "defending Israeli lives" (or security or protection), the Iron Dome may simultaneously be associated with other values by other actors, e.g., Hamas.

Here again, we can see the problem with the idea of embedded values and thinking of AI as merely computational artifacts. If we take van de Poel's approach, discussed earlier, we might imagine the Israeli military deciding that quicker updates of the code of the Iron Dome are needed because it is not effective against the new tactics of Hamas. This seems to fit van de Poel's feedback loop model: the Israelis deploy the AI, Hamas organizes missile attacks that reveal the AI's vulnerabilities, the Israeli military updates the AI to counter those attacks, and so on, in a (literal) arms race of ongoing enhancements to AI technology. In this description, the focus seems to be on code and from a computational and operational perspective, van de Poel's model might be a good way to describe what is happening with the Iron Dome as it is designed and redesigned. However, this perspective only includes the AI designers, who obviously play a fundamental role in the creation of the Iron Dome itself but leaves out many other social factors and actors that lead to values being attributed to system. In van de Poel's account, something like "defending Israeli lives" might be considered the intended value, in that the designers intend to create an AI artifact that embodies such value. Imagine that, after its deployment, the Iron Dome is successfully fending off all missile attacks from Hamas. Again, in van de Poel's model, the value would be said to be "realized" in the Iron Dome, because not only is it intended by the artifact's designers and "embedded" in the artifact, but it is also achieved in the relevant deployment and use.

What, then, happens to that embedded value when Hamas soldiers come up with a new missile attack tactic that succeeds in breaching the Iron Dome and causes victims among the Israeli people? Was the value "*dis*embedded"? Did the opposite value of "harming Israeli lives" become "embedded" or "realized" in the Iron Dome? Was Hamas the actor that did the "disembedding" of the value or the embedding of the opposite value? To adopt van de Poel's terminology: what happens to the embeddedness of a value that was previously embedded and also realized, when not only is it not realized anymore but we also have effects that one would associate with the very opposite value?

We see here that a change in the values attributed to the Iron Dome was not the result of a change in design or coding but a change in the behavior of targets of the system. Here we see that values are not 'in' AI, they are attributed to AI by humans because of the effects of the AI and the way it operates in relation to things that people care about. Values are not in the code and strings of code do not contain values. Code operates in systems with social relations that produce effects, both physical and social, and people associate those effects with conditions that they value in particular ways.

So again, now in the Iron Dome AI, we see that the relationship between values and AI is best understood if we think about AI as a sociotechnical system, that is, a combination of computational artifacts and social relations. Whatever values are attributed to the Iron Dome, they can only be understood as such by treating the Iron Dome as computational artifacts *together with* a set of social relations, social ideas, social interests and arrangements.

One caveat is in order here. Although we argue that the relationship between values and AI can best be understood using the concept of AI as sociotechnical systems, we have shied away from saying that values can *never* be attributed to AI understood purely as computational artifacts. This is because humans attribute values to all kinds of things using language in a variety of ways, descriptively, metaphorically, performatively, fancifully, etc. Any given artifact can be described in a multiplicity (if not an infinite number) of ways. So, it is possible for groups of people such as computer scientists to agree, and they often do agree, to refer to a particular line of code as an instantiation of a value. They may concur, for example, that a specific part of the Iron Dome software is "defending Israeli lives" because those lines of code determine the position of Hamas' missiles. If such conceptualization helps the team of designers and engineers work better, so be it. However, for those who are not part of the team—especially those who are discussing AI in the public sphere—to speak this way is to mislead. The sociotechnical concept of AI provides a more accurate picture of how AI can have a relationship to values. Only this concept recognizes that AI includes people and it is people who entertain notions of values.

5 Conclusion

Our argument is, then, that the two concepts of AI lead to a good deal of confusion. The values related to AI can only adequately be understood using the concept of AI as sociotechnical systems. Confusion results when people switch back and forth from one concept to another in the same analysis. In conflating the two concepts of AI in this way, they are misled into thinking that values are simply in the computational components of AI. This leads to notions of "embedded values" and values being "embodied" in AI. Unfortunately, this way of thinking suggests that the problems associated with AI and values can be solved by greater attention to the code and through re-coding. To be sure, some of the problems with AI can be addressed through re-design but the re-design will have to involve much more than re-coding.

Recognizing the difference between the two concepts is particularly important when attempts are made at formulating prescriptive principles on how to build AI systems that afford this or that value. The "AI Act" recently proposed by the European Commission, for instance, features a good deal of discourse on AI and values (European Commission 2021): AI systems with "unacceptable risk" will be considered a threat to people and will be banned, including cognitive behavioral manipulation of people or specific vulnerable groups, like voice-activated toys that encourage "dangerous behaviour" in children. Concepts like "unacceptable", "manipulation", and "dangerous" are not fixed, static, or immutable. They are up for debate, amenable to change, context-dependent, culturally variable…in one word: they are social.

To be sure, ensuring that the adoption of AI systems does not work against but facilitates the things that people value is a daunting task. What is clear from the preceding is that the challenge is not just one of computational design and, for this reason, it should not be left only to engineers and computer scientists. If we want AI systems that promote rather than undermine important social values, attention needs to be given to the social relations, arrangements, and concepts that will constitute those systems.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Funding Open access funding provided by Università degli studi di Bergamo within the CRUI-CARE Agreement.

Data availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Arkin RC (2009) Governing lethal behavior in autonomous robots. Chapman & Hall, Boca Raton

Baxter G, Sommerville I (2011) Socio-technical systems: from design methods to systems engineering. Interact Comput 23(1):4–17

Borning A, Muller M. (2012) Next steps for value sensitive design. In: *CHI'12 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1125–1134

Braithwaite V, Blamey R (1998) Consensus, stability and meaning in abstract social values. Aust J Polit Sci 33(3):363–380

Branscomb L, Auerswald PE. (2002) *Between Invention and Innovation. An Analysis of Funding for Early-Stage Technology Development.* Technical Report 02–841. National Institute of Standards and Technology.

Chatila R, Havens JC (2019) The IEEE global initiative on ethics of autonomous and intelligent systems. In: Ferreira MIA et al (eds) Robotics and Well-Being. Springer, Cham, pp 11–16

Cieciuch J, Schwartz SH (2018) Values and the human being. In: van Zomeren M, Dovidio JF (eds) The oxford handbook of the human essence. Oxford University Press, Oxford, pp 219–231

Coeckelbergh M (2020) Challenges for policymakers. In: Coeckelbergh M (ed) AI ethics. MIT Press, Cambridge, pp 167–181

de Boer B (2021) Explaining multistability: postphenomenology and affordances of technologies. AI Soc. https://doi.org/10.1007/s00146-021-01272-3

Etzioni A, Etzioni O (2016) Designing AI systems that obey our laws and values. Commun ACM 59(9):29–31

European Commission (2021) Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

Finch J, Geiger S, Reid E (2017) Captured by technology? How material agency sustains interaction between regulators and industry actors. Res Policy 46(1):160–170

Friedman B, Hendry DG (2019) Value sensitive design: Shaping technology with moral imagination. MIT Press, Cambridge

Garrety K, Badham R (2004) User-centered design and the normative politics of technology. Sci Technol Human Values 29(2):191–212

Gentile MC (2010) Giving voice to values. Yale University Press, New Haven

Greene J, Rossi F, Tasioulas J, Venable KB, Williams B (2016) Embedding ethical principles in collective decision support systems. Proc Thirtieth AAAI Conf Artif Intell 30:4147–4151

Hofstetter R, Dahl DW, Aryobsei S, Herrmann A (2021) Constraining ideas: how seeing ideas of others harms creativity in open innovation. J Mark Res 58(1):95–114

Hong HY, Sullivan FR (2009) Towards an idea-centered, principle-based design approach to support learning as knowledge creation. Educ Tech Res Dev 57(5):613–627

Ihde D (1999) Technology and prognostic predicaments. AI Soc 13:44–51

Jasanoff S (ed) (2004) States of knowledge: the co-production of science and the social order. Routledge, London

Johnson DG, Verdicchio M (2017) Reframing AI discourse. Mind Mach 27(4):575–590

Kim TW, Hooker J, Donaldson T (2021) Taking principles seriously: a hybrid approach to value alignment in artificial intelligence. J Artif Intell Res 70:871–890

Klenk M (2021) How do technological artefacts embody moral values? Philos Technol 34:525–544

Kop M (2020) The right to process data for machine learning purposes in the EU. SSRN J. https://doi.org/10.2139/ssrn.3653537

Lambrecht A, Tucker C (2019) Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. Manag Sci 65(7):2966–2981

Lanzeni D, Pink S (2021) Digital material value: designing emerging technologies. New Media Soc 23(4):766–779

Latour B (2005) Reassembling the social: an introduction to actor-network-theory. Oxford University Press, Oxford

Le Dantec CA, Poole ES, Wyche SP (2009) Values as lived experience: evolving value sensitive design in support of value discovery. In *CHI'09 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1141–1150

Li AY (2020) Performance funding policy impacts on STEM degree attainment. Educ Policy 34(2):312–349

Little D (2000) Explaining large-scale historical change. Philos Soc Sci 30(1):89–112

Misa TJ (1988) How machines make history, and how historians (and others) help them to do so. Sci Technol Human Values 13(3–4):308–331

Nickel PJ, Kudina O, van de Poel I (2022) Moral uncertainty in technomoral change: bridging the explanatory gap. Perspect Sci 30(2):260–283

Oudshoorn N, Pinch T (2003) How users matter: the co-construction of users and technology (inside technology). The MIT Press, Cambridge

Paz MJ, Rísquez M, Ruiz-Gálvez ME (2021) Inter-firm power relations and working conditions under new production models. *The Economic and Labour Relations Review* 1–20

Rokeach M (1973) The nature of human values. Free Press, Mumbai

Roy R, Goatman M, Khangura K (2009) User-centric design and Kansei engineering. CIRP J Manuf Sci Technol 1(3):172–178

Salo-Pöntinen H (2021) AI ethics-critical reflections on embedding ethical frameworks in AI technology. In: Rauterberg M (ed) Culture and computing: design thinking and cultural computing. 9th International Conference, C&C 2021, held as part of the 23rd HCI international conference, HCII 2021, virtual event, July 24–29, 2021, proceedings, Part II, vol 12795. Springer, Cham, pp 311–329

Schwartz SH (1994) Are there universal aspects in the structure and contents of human values? J Soc Issues 50(4):19–45

Scott JF (1971) Internalization of norms: a sociological theory of moral commitment. Prentice-Hall, Hoboken

Shahriari K, Shahriari M (2017) IEEE standard review—Ethically aligned design: a vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In: Shahriari K, Shahriari M (eds) 2017 IEEE Canada International Humanitarian Technology Conference (IHTC). IEEE, Piscataway, pp 197–201

Swierstra T, Stemerding D, Boenink M (2009) Exploring techno-moral change: the case of the obesitypill. In: Sollie P, Düwell M (eds) Evaluating new technologies: methodological problems for the ethical assessment of technology developments. Springer, Dordrecht, pp 119–138

Tollon F (2022) Artifacts and affordances: from designed properties to possibilities for action. AI Soc 37:239–248

Torresen J (2018) A review of future and ethical perspectives of robotics and AI. Front Robot AI 4:75

Umbrello S (2019) Beneficial Artificial Intelligence coordination by means of a value sensitive design approach. Big Data Cogn Comput 3(1):5

Umbrello S, Capasso M, Balistreri M, Pirni A, Merenda F (2021) Value sensitive design to achieve the UN SDGs with AI: a case of elderly care robots. Mind Mach 31(3):395–419

van de Poel I (2020) Embedding values in Artificial Intelligence (AI) systems. Mind Mach 30(3):385–409

van den Hoven J (2013) Value Sensitive Design and Responsible Innovation. In: Owen R, Bessant J, Heintz M (eds) Responsible innovation: managing the responsible emergence of science and innovation in society. Wiley, Hoboken, pp 75–83

van der Merwe J (2021) Iron Dome Shows AI's Risks and Rewards. CEPA.org, June 1 2021. https://cepa.org/iron-dome-shows-ais-risks-and-rewards/

van de Poel I, Kroes P (2014) Can technology embody values? In: Kroes P, Verbeek PP (eds) The moral status of technical artifacts. Springer, Dordrecht, pp 103–124

Walker W (2000) Entrapment in large technology systems: institutional commitment and power relations. Res Policy 29(7–8):833–846

Whitford AB, Tucker JA (2009) Technology and the evolution of the regulatory state. Comp Pol Stud 42(12):1567–1590

Wolf MJ, Miller KW, Grodzinsky FS (2017) Why we should have seen that coming: comments on Microsoft's Tay "experiment", and wider implications. ORBIT J 1(2):1–12

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