Discover Artificial Intelligence

Perspective

Different approaches to the moral status of AI: a comparative analysis of paradigmatic trends in Science and Technology Studies

Roberto Redaelli[1]

Received: 16 June 2023 / Accepted: 6 July 2023
Published online: 17 July 2023

Abstract
The exponential progress of AI systems today compels scientists and philosophers to redefine their conceptual frameworks to better understand the nature of these new technologies and their moral status. Among the various theories that are used to respond to the challenges posed by intelligent systems are instrumentalism, Socio-technical Systems Theory (STST) and Mediation Theory (MT), all widely adopted in the field of Science and Technology Studies (STS). This paper intends to present the main features of these theories and provide a comparative analysis of them in order to assess their contribution to the process of understanding the moral status of artificial intelligence. Our investigation intends to show how (1) instrumentalism is inadequate to account for the moral status of AI, (2) STST, while helping to highlight the link between AI, society and morality, lends itself to the criticism of anthropocentrism, (3) MT in its Latourian version has the merit of highlighting the active character of technological artefacts and thus of artificial intelligence in the moral sphere. However, the principle of symmetry it proposes poses the problem of the de-accountability of the human agent. (4) MT in its postphenomenological form seems to partially resolve the problem of moral responsibility, but the opacity of the terminology it employs exposes it to various criticisms. In light of these results, we intend to show how an understanding of the moral status of intelligent systems cannot be based on the diametrically opposed positions that consider technologies either morally neutral or else moral agents similar to humans, whereas particularly useful elements can be found in STST and in postphenomenological MT.

Keywords Ethics of artificial intelligence · Philosophy of technology · Moral philosophy · Artificial moral agents · Moral standing of AI

1 Introduction

The pervasiveness with which intelligent systems currently shape our society raises increasingly pressing ethical and legal questions. The use of robots in care and education, the spread of decision support systems in the legal and medical fields, and the adoption of increasingly powerful communication tools are radically redefining our patterns of behaviour and thought.

This epochal change has become the focus of scientists and philosophers who seek new conceptual frameworks to understand the true nature of AI. Moreover, recent advances in AI have highlighted how the merely instrumentalist common view is no longer sufficient to understand the status of new technologies and the moral role they play in our society. It is precisely the clarification of the ethical impact of intelligent systems and their moral standing that constitutes one of the most pressing challenges in the field of STS today, a challenge that has given rise to a wide-ranging debate in recent

✉ Roberto Redaelli, roberto.redaelli@unimi.it | [1]Department of Philosophy "Piero Martinetti", University of Milan, Milan, Italy.

years. We intend here to analyse three paradigmatic positions of this multifaceted debate: the Value Neutrality Thesis (VNT), the Socio-technical Systems Theory (STST) and the Mediation Theory (MT),[1] with the aim of highlighting their limitations and strengths. With the comparative analysis of these theories, we intend to show how an understanding of the moral status of intelligent systems, endowed with varying degrees of autonomy, interaction and adaptability, cannot be based on the diametrically opposed positions that consider technologies either as morally neutral or else moral agents on a par with humans, whereas particularly useful elements are to be found in STST and postphenomenology-based MT. In fact, these theories, albeit within certain limits, mitigate certain problems related to instrumentalism and Latour's mediation theory. More precisely, this investigation intends to show how (1) instrumentalism is inadequate to account for the moral status of AI. (2) While STST helps to bring to the forefront the link between AI, society and morality, it also exposes itself to the criticism of anthropocentrism. Due to its anthropocentrism STST is accused of falling into a more refined form of instrumentalism than the traditional one. (3) MT in the version proposed by Latour has the merit of highlighting the active character of technologies in the social sphere and thus of artificial intelligence. However, the principle of symmetry it proposes raises the problem of the deresponsabilization of the human agent. (4) MT in its postphenomenological version seems to partially resolve the problem of moral responsibility, but the opacity of the terminology it employs exposes it to numerous criticisms.

2 Strong and weak value neutrality thesis (VNT)

Much of our everyday understanding of technology is based on the assumption that a technological object is merely a tool that is useful for achieving a certain purpose. This so-called instrumentalist view therefore considers all technological objects—be they a hammer or a humanoid robot—as tools with no moral value. But the uses that are made of such technologies are subject to moral evaluation. In this sense, we speak of a dual use of technological artefacts: they can be used for morally good or bad ends. We consider two variants of this theory here: one strong, one weak.

A strong version of the instrumentalist position is supported today by Joseph Pitt, for whom "technological artefacts do not have, have embedded in them, or contain values" [1]. In this radical form defined as the Value Neutrality Thesis (VNT), instrumentalism rules out the idea that artefacts can incorporate values. As proof of this thesis, Pitt cites two famous examples. He invokes the slogan of the National Rifle Association of America: "Guns don't kill people, people kill". In this case, in Pitt's analysis, it is the use of the firearm and not the firearm itself that has a moral value. In fact, the firearm does not exert any coercive force that drives the individual to commit a crime, but it is the holder of the firearm who decides how to use it.

The second example is taken from the famous case of the Long Island overpasses designed by architect Robert Moses. These overpasses were deliberately designed by Moses, at the beginning of the last century, in such a way that no buses could pass under them. This prevented black people and anyone who could not afford a car from reaching the well-known beaches that were accessible to white people and the upper-middle classes. According to Langdon Winner's analysis of these overpasses [2], these artefacts, designed too low for public transport vehicles to pass under them, were promoters of elitist or racial discrimination. In other words, they incorporated a political order. In contrast to Winner's analysis, Pitt argues that in no way can these overpasses be said to incorporate values: the only values involved here are those of Moses—it cannot be claimed that overpasses *have* such values. Therefore, for Pitt, although the construction and design of technologies involves numerous value-laden decisions, the conclusion that the artefacts themselves are value-laden cannot be drawn.

A weak version of the instrumentalist thesis is expressed by Peterson and Spahn [3]. What seems to distinguish the strong thesis from the weak one is that, although they agree that (1) technological artefacts are not considered moral agents and (2) they are not responsible for their effects, they can sometimes "affect the moral evaluation of general actions" [3, p. 412]. For example, according to the authors, from a consequentialist perspective, *the presence or absence*

[1] These theories appear paradigmatic because they well represent three different tendencies dominating the debate on the moral status of AI. Instrumentalism exemplifies all positions that regard AI as morally neutral. At the opposite extreme of this theory we can place the mediation theory proposed by Latour, which, by equating human and non-human actors, represents in its most radical form all positions that assign to technologies in general a full moral agency on equal terms with human agency. Postphenomenological mediation theory and STST represent two intermediate positions, the former being post-humanist, the latter anthropocentric. These are complemented by numerous theories that can be located to varying degrees within the coordinates represented by these positions, and which we will not address here.

of a bomb in the suitcase of a terrorist who intends to kill a million people by pressing a button changes the evaluation of the action itself. Put another way, if there were no bomb, the moral evaluation of *pushing the button* would be radically different.

In contrast to both versions of instrumentalism—we emphasise that the weak one concedes that while technology is neutral, it has an effect on the *moral evaluation* of action—the problem of the incorporation of values and disvalues peculiar to intelligent systems should be discussed here. In particular, the issue of biases affecting intelligent systems is currently a common theme in the field of STS. Indeed, numerous studies have demonstrated how intelligent systems, in addition to being able to perform the function of artificial moral advisor [4], have the capacity to embody values and biases [5, 6]. More precisely, algorithms, when they are based on incomplete or non-representative data, can promote different types of discrimination [7]. One example that has become famous is that of judges who in some legal systems use algorithms to determine the risk of recidivism. As in the Compas case [8], there may be biases that affect the decision support tools used by judges, since their decisions are based on previous resolutions, according to a bottom-up statistical approach.

To deal with the problem of incorporating disvalues, more and more attention has been paid in recent years to the design phase of intelligent systems [9] and in particular to the data to be used in training them. Both the design phase and the choice of data to be used are, in fact, two decisive moments in the management of the ethical impact of intelligent systems. Although not all consequences of technologies are predictable, it is still possible to design technologies in such a way as to deliberately convey certain values (e.g. health, safety, privacy) and avoid the dissemination of disvalues (e.g. racial and gender discrimination). In this sense, technology cannot be seen as a mere morally neutral tool in the hands of users, but rather it takes on clear moral meaning right from the design phase. This meaning depends primarily on the social function for which it is supposed to be used and the way it is designed, i.e. the values it is intended to embody.

This moral meaning is even better exemplified in certain intelligent systems that aim to support medical decisions by assessing their ethical impact. A good example is offered today by MedEthEx [10]. Designed by Michael Anderson, Susan Leigh Anderson and Chris Armen, this prototype of decision support aims to assist physicians in resolving ethical dilemmas that may arise in medical practice. For example, in the case of a patient who refuses a transfusion for religious reasons and thereby puts his life at risk, the following question arises: should the doctor try to change the patient's mind (impinging upon the patient's autonomy) or accept his choice (violating his/her duty to provide the most beneficent care)? It is clear in this example that the doctor's decision support system is not only not morally neutral, but is intended to contribute decisively to the resolution of certain ethical questions.

Because of the ability of AI systems to incorporate (intentionally or unintentionally) values and disvalues, and also their ability to offer decision-making support in the presence of increasingly pressing moral demands, it is not possible to affirm that such systems have the status of mere morally neutral instruments. This limitation can also be seen in what is referred to nowadays as Instrumentalism 2.0 [11]. This expression is meant to indicate a renewed application of instrumentalism to new technologies, in which the instrumental thesis is, so to speak, radicalised: robots and new technologies are reduced to mere slaves [12] at our service. They are our property, just as a toaster and a hammer are. In regard to the proposal put forward by so-called Instrumentalism 2.0, it is sufficient to observe that this new version of instrumentalism does not offer any new elements that could free it from the same limitations as the VNT presented by Pitt. However, we must conclude by observing that, despite the inability of this theory to account for systems with high levels of autonomy, interaction and adaptability, it can still be considered valid for very simple tools and technologies, for which their use rather than their design is most important.

3 AI as socio-technical system

A different perspective that can lead us closer to understanding the moral status of intelligent systems is what is known as Socio-technical Systems Theory (STST). Unlike the VNT, this theory tends initially to emphasise the dual nature of the technological artefact, constituted by a material and a social element. Indeed, the artefact is characterised by a physical composition and the practical function for which it was designed. To fulfil its social function, linked to human goals that mostly possess moral significance [13], the artefact is embedded in a use plan. This term is understood as "a plan that describes how an artefact should be used to achieve certain goals or to fulfil its function" [14, p. 391]. Therefore, in the design phase, engineers and developers are responsible not only for the form the product is to take, but also for its possible uses. In this sense, engineers and developers can design artefacts in such a way that they are capable of moral behavioural influencing, facilitating good practices and/or avoiding bad uses.

By further broadening the perspective from which to investigate the technological object, STST recognises not only that the artefact has a hybrid (material and social) nature, but that it forms part of what is defined in terms of a socio-technical system. By this expression is meant a system whose functioning depends both on the technical element and on human behaviour and institutions understood as rules to be followed [6, 15]. Examples of such types of systems are a company, or the field of civil aviation. In the latter case, the various parts of the system, i.e. the technological apparatus (e.g. aircraft, technological checkpoint systems such as scanners, etc.) the human agents (on-board and ground personnel) and the laws and regulations in the field of aviation enable the proper functioning of the system itself, when each part performs its task properly.

Returning to AI systems, according to Socio-technical Systems Theory, what distinguishes a traditional socio-technical system from an AI system is the presence in AI systems of certain artificial components that replace or accompany human agents. More precisely, in AI systems, artificial agents replace or work alongside human agents, while institutions, understood as social rules, are 'translated' into technical norms. In this way, the introduction of artificial agents and technical rules into the system entails a redefinition in causal terms of certain parts of the system itself, which in traditional Socio-technical Systems Theory are represented by the intentionality of human agents and institutions.

For our purposes, it is important to note here that AI systems, in this perspective, are characterised by properties such as autonomy, interactivity, and adaptivity [16], which are also possessed by human agents [6]. However, despite the fact that such properties are shared by humans and AI systems, STST proponents do not tend to attribute to the latter a moral agency equal to that of humans, since such artificial agents lack other exquisitely human characteristics such as consciousness, intention to act, and freedom.

An example of the use of this conceptual framework is offered by Deborah G. Johnson [17] in the field of computer ethics. The author recognises that computers are parts of socio-technical systems, and thus attributes to them the moral status of *moral entities* precisely by virtue of their belonging to the system. In fact, although computers cannot be considered moral agents since they lack mental states and intentions to act, they possess moral significance since they are linked to the human intentionality that brought them into being, to the social practices in which they are employed, and to certain knowledge systems. In fact, they incorporate the intentionality of users and designers into their own intentionality, which is related to their functionality, thus forming a triad of intentionality that is at work in technologies. For this reason, Johnson observes that "no matter how independently, automatically, and interactively computer systems of the future behave, they will be the products (direct or indirect) of human behavior, human social institutions, and human decision" [17, p. 197].

This position, which has the undisputed merit of clarifying the moral significance of computers and intelligent systems, nevertheless seems to incur a criticism that once again relates to instrumentalism. Johnson's position appears to be spoiled by a certain anthropocentrism: it would reduce intelligent systems to a mere *extension* of the human subject, without attributing to them an autonomous character and a moral impact (positive or negative) that many times, in the case of AI systems, cannot be reduced only to the one intended by designers and users. In this sense, for Gunkel, "although Johnson's 'triad of intentionality' is more complex than the standard instrumentalist position, it still proceeds from and protects a fundamental investment in human exceptionalism. Despite considerable promise to reframe the debate, Johnson new paradigm does not look much different from the one it was designed to replace. Human beings are still and without question the only legitimate moral agents" [18].

Although dismissing Johnson's position to a more complex form of instrumentalism than the classical one might seem too radical, it must be emphasised that the behaviour of some technologies such as machine learning, which may lead to AI technologies to 'disembody' the values embedded in them [19], are difficult to explain in terms of the human–machine relationship. In other words, not all aspects and operations of technologies can be linked exclusively to the intentions of the human who put them in place. In this sense, precisely on the basis of the example of machine learning, it is difficult to share Johnson's idea that.

> "even when it learns, it learns as it was programmed to learn. […] The fact that the designer and user do not know precisely what the artifact does makes no difference here. It simply means that the designer – in creating the program – and the user – in using the program – are engaging in risky behavior. They are facilitating and initiating actions that they may not fully understand, actions with consequences that they cannot foresee. The designer and users of such systems should be careful about the intentionality and efficacy they put into the world. […] When humans act with artifacts, their actions are constituted by their own intentionality and efficacy as well as the intentionality and efficacy of the artifact which in turn has been constituted by the intentionality and efficacy of the artifact designer" [17, pp. 203–204].

In light of the critique by Gunkel and others, considering the unintended consequences of technologies only in terms of risky behaviour on the part of humans seems to disregard the active character of technologies and the emergent abilities they may exhibit.[2] In this sense, while the position of Johnson and STS has the undoubted merit of highlighting the moral significance of technologies in relation to humans, it does not seem to be able to offer an exhaustive explanation of intelligent systems, whose autonomy, adaptability and interactivity is often not imputable to the intentionality of designers and users. The fact that machines endowed with the capacity to learn develop their own characteristics seems therefore to require a conceptual framework that recognises a certain moral agency and a congenital pre-intentional nature,[3] and considers them more than mere tools or extensions of the human subject, even without attributing to them any understanding, capacity for reasoning or moral action equal to that of humans.

4 Mediation theory: from Latour to postphenomenology

A radical re-evaluation of the ontological status of technological artefacts, with obvious repercussions on the clarification of the moral status of intelligent systems, is found in the theory of mediation inaugurated by Bruno Latour [20–22], later re-elaborated in postphenomenological terms by Peter-Paul Verbeek.

With regard to Actor-Network theory, we are interested here in highlighting two points of particular importance for our understanding of the moral status of artificial agents. Firstly, Latour is credited with drawing attention to the active character of artefacts using the principle of symmetry. Technological artefacts are not merely the product of social networks, nor are they morally neutral instruments in our hands. Technologies are part of networks composed of human and non-human actors, within which these actors occupy different yet symmetrical positions. According to Latour, their identities are not fixed, but depend on the position they occupy in the network, just as—we would add—their moral role, too, depends on their position.

With this redefinition of the ontological status of technologies, and more generally of what is non-human, Latour intends to overcome modern subject-object, culture-nature dichotomies in order to highlight how our world is formed by hybrid collectives. In such collectives, humans and non-humans are constituted reciprocally, or rather, man and technology form a new subject, a new *actant*.

An example of the explanatory model proposed by this theory is once again provided by firearms. The firearm is not, in Latour's eyes, a merely neutral instrument, as it is for the VNT. The person who possesses a firearm forms a new subject that ineluctably modifies both the person who possesses it and the firearm itself [21]. More precisely, possession of a firearm transforms the desires, skills, and possibilities of its possessor, and also transforms the firearm itself, which in the hands of the individual has a new identity: it can kill, unlike, for example, the weapon placed in a cabinet in an army museum. In this sense, we can say that the weapon, or the artefact, is a mediator, to which a program of action is delegated whereby moral action is distributed between human and non-human actors. This delegation is particularly evident in the case of intelligent systems that are involved in networks of relationships and cannot be reduced to a mere extension of the human subject, as is the case within the framework proposed by STST. Intelligent systems and humans constitute each other and it is within the collectives of human and non-human actants that the moral sphere takes shape.

Now, despite the broad impact of this theory on contemporary scenarios, there is a clear problem from a moral point of view that applies especially to intelligent systems. If one considers human and non-human subjects according to the symmetrical model, the problem arises of the responsibility to be assigned to non-human subjects. Who is responsible for damage caused, for example, by a self-driving car? Can an intelligent system be held responsible for its actions? Given that Latour proposes distributing responsibility among the various actants, the question of how such a distribution is possible between humans and systems with artificial intelligence becomes increasingly urgent. Despite a number of solutions proposed today (among the various possibilities of redistribution of responsibility, see, e.g., [23]), this question still remains open.

[2] In this respect, one only has to think of the emerging abilities of large language models analysed using the so-called BIG-Bench benchmark, whereby these models present 'quantitative improvement and new qualitative capabilities with increasing scale'.

[3] By pre-intentional nature we refer here, taking up the terminology presented by Di Martino in *Viventi umani e non umani*, Milano, Cortina, 2017, to the fact that the effects of AI-enabled technologies go beyond the intentionality of the designers and users and have feedback effects on the human him/herself.

This is compounded by a second problem. As noted by Verbeek, Latour's position reduces human and non-human actants to parts of collectives and thus risks losing sight of the mediating role that technology plays in the relationship between humans and the world. In fact, while Latour's position certainly has the merit of emphasising the active character of technologies from an outside perspective, it reduces the connections between entities to mere associations [24]. Because of this reduction, it does not seem to provide an adequately nuanced look at these same connections. On the other hand, in Verbeek's view, the postphenomenology inaugurated by Don Ihde [25–27], which provides an analysis of the human-technological relationship from the standpoint of human experience, from an internal perspective, so to speak, seems to succeed in this endeavour. Postphenomenology, in fact, and here we refer to the version proposed by Verbeek, performs a twofold operation with regard to Latour's theory of mediation: on the one hand, it rejects the symmetry between human and non-human actants, and on the other hand, it accepts the notion of mediation. With this gesture, it succeeds in a single step in accounting for the moral role played by technologies, without, however, leading to a deresponsabilization of the human subject. Technologies are moral mediators, that is, they mediate our relationship to the world, they play an active role in our experience of the world, but without becoming themselves responsible. If anything, Verbeek notes, it is necessary to grasp the role that technologies play in shaping our responsibility. The introduction of a new technology (e.g. technologies for prenatal diagnoses [28]) broadens our possibilities for action and thus the domain of our responsibility [29]. In fact, although technologies direct human actions and decisions, it is the human who must decide which options to pursue.

Returning to the question of artificial intelligence, one could observe that, according to the postphenomenological perspective, artificial intelligence is a moral mediator of a special kind, i.e. it is endowed with autonomy, interactivity and adaptability. Such a moral mediator modifies our experience of the world, our action, gaining the moral status of an agent, without for that reason requiring responsibility and intentions to act. Now, although it lacks intentions to act, according to the postphenomenological approach, AI is endowed with intentionality arising from the interaction with the person who designs and uses it, but AI is not completely reducible to human intentionality. In this sense, Verbeek rightly speaks of emergent forms of mediation, in which all the emergent (moral) abilities with which intelligent systems are endowed could converge.

An example of the explanatory power of this approach can be offered by a brief consideration of LLMs that have revolutionised natural language processing (NLP) today. These deep learning models present emergent abilities that are not present in smaller models. Such abilities may have a clear moral significance, as in the case of identifying offensive content in paragraphs of Hinglish (a combination of Hindi and English) or in the so-called capacity for moral self-correction [30]. In this case, AI is not merely a morally neutral tool, nor is its intentionality reducible to that of users and designers, just as its emergent abilities are by no means definable in terms of human risky behaviour.

In this way, postphenomenological theory, thanks to the notion of moral mediation, seems to offer several useful elements for understanding the moral status of AI systems. It highlights the autonomous character of such systems and is able to account for the more than instrumental character of AI, even though it does not assign it a moral agency equal to that of humans. Despite these strengths, postphenomenological mediation theory has limitations to which several critics have drawn attention. Among these limitations is the problem of terminology, which seems to undermine the foundations of the entire postphenomenological construct (On the problem of the terminology used by Verbeek, see [31]). In fact, Verbeek employs expressions such as the 'morality of things' or the 'moral agent of technology', with which he seems to assign to technologies in general a moral status equal to that of humans. This would reintroduce the problem of the moral responsibility of artefacts, as well as that of intentions to act, which intelligent systems lack. In order to avoid such problems, more precise control of the terminology used would lead to a better understanding of the strengths of this theory [32].

5 Conclusions

The analysis we have undertaken of the VNT, STST, and MT (in the Latourian and postphenomenological versions) has led to the following results on the question of the moral status of artificial intelligence:

a) Although the conceptual framework proposed by the VNT can be used for technologies simpler than AI, it fails to take into account phenomena such as the incorporation of values and disvalues that arise in intelligent systems.

b) STST has the merit of bringing the ethical significance of AI to the forefront through the notion of a moral entity. With this notion, STST attributes a certain intentionality to intelligent systems, not understood as an intention to act, but as efficacy and functionality. The limitation of this position, at least in Johnson's version, is that it does not seem to take into account the emerging moral properties of AI and reduces them instead to human risky behaviour. In fact, the ever-increasing operational autonomy of intelligent systems requires a framework that is able to account for AI's emerging abilities (moral and non-moral) which cannot be limited to the intentionality of designers and users. STST is also accused by some critics of anthropomorphism and considered only as a more complex form of instrumentalism.

c) The MT developed by Latour has the merit of highlighting the active character of technologies within so-called collectives. The equating of human and non-human actants proposed by this theory, however, raises the problem of the moral responsibility to be attributed to technologies. Added to this is the equalisation of human and non-human actants, reduced to mere parts of the same collectives, with the consequent blurring of the different types of technological mediation present in the human-world relationship.

d) The postphenomenological theory of mediation has the advantage of clearly emphasising the mediating role played by technologies, without attributing any form of moral responsibility to them. This theory also seems to adequately consider emerging forms of moral mediation. Despite these merits, we have observed that this approach is not without certain limitations. Among these shortcomings, we have highlighted a certain opacity in the terminology employed by Verbeek. He makes use of expressions such as 'morality of things' or 'moral agent of technology' that can undermine the entire postphenomenological framework.

From these results, it can be observed that the problem of the moral status of technological artefacts, while remaining an open question, finds useful elements for its resolution in the theories analysed here. These elements, although they cannot form a well-defined mosaic, are valuable paths that can be explored in further investigations into the moral status of artificial intelligence. They also show, by virtue of the comparative analysis we have conducted, that an understanding of the moral status of AI cannot be based on the diametrically opposed positions that regard technologies either as morally neutral or as moral agents on a par with humans. The first position takes an overly simplistic view of technologies that cannot account for the recent developments in AI and the ethical implications of these developments. The second position does not seem to take into account the differences between humans and AI systems. On these differences both STST and postphenomenology converge: intelligent systems, although endowed with intentionality (understood as directionality by Verbeek and Ihde, and as efficacy and functionality by Johnson) are devoid of intentions to act and mental states to which the sphere of freedom is linked.

Beyond the convergences and differences that exist between the intermediate positions of STST and the mediation theory of the phenomenological type, these positions overcome, as we have shown, some of the problems that encumber Latour's theory and the instrumentalist thesis, and thus provide some useful tools in the complex process underway to understand the moral status of AI.

Author contributions RR wrote the manuscript. The author read and approved the final manuscript.

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

Data availability No datasets were generated because the paper takes a theoretical approach.

Declarations

Competing interests The authors declare no competing interests.

References

1. Pitt JC. "Guns don't kill, people kill"; values in and/or around technologies. In: Kroes P, Verbeek PP, editors. The moral status of technical artefacts. Dordrecht: Springer; 2014. p. 89–101. https://doi.org/10.1007/978-94-007-7914-3_6.
2. Winner L. The whale and the reactor: a search for limits in an age of high technology. 2nd ed. Chicago: University of Chicago Press; 1986.
3. Peterson M, Spahn A. Can technological artefacts be moral agents? Sci Eng Ethics. 2011;17:411–24. https://doi.org/10.1007/s11948-010-9241-3.
4. Giubilini A, Savulescu J. The artificial moral advisor. The "ideal observer" meets artificial intelligence. Philos Technol. 2018;31:169–88. https://doi.org/10.1007/s13347-017-0285-z.
5. Flanagan M, Howe D, Nissenbaum H. Embodying values in technology: theory and practice. In: van den Hoven J, Weckert J, editors. Information technology and moral philosophy. Cambridge: Cambridge University Press; 2008. p. 322–53.
6. van de Poel I. Embedding values in artificial intelligence (AI) systems. Mind Mach. 2020;30:385–409. https://doi.org/10.1007/s11023-020-09537-4.
7. Kirkpatrick K. Battling algorithmic bias: how do we ensure algorithms treat us fairly? Commun ACM. 2016;59:16–7. https://doi.org/10.1145/2983270.
8. Brennan T, Dieterich W, Ehret B. Evaluating the predictive validity of the COMPAS risk and needs assessment system. Crim Justice Behav. 2009;36:21–40. https://doi.org/10.1177/0093854808326545.
9. Umbrello S, van de Poel I. Mapping value sensitive design onto AI for social good principles. AI Ethics. 2021;1:283–96. https://doi.org/10.1007/s43681-021-00038-3.
10. Anderson M, Anderson S, Armen C. MedEthEx: toward a medical ethics advisor. Papers from the 2005 AAAI Fall Symposium.
11. Gunkel DJ. Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf Technol. 2020;22:307–20. https://doi.org/10.1007/s10676-017-9428-2.
12. Bryson JJ. Robots should be slaves. In: Wilks Y, editor. Close engagements with artificial companions: key social, psychological, ethical and design issues. Amsterdam: John Benjamins; 2010. p. 63–74.
13. Vermaas P, Kroes P, van de Poel I, Franssen M, Houkes W. A philosophy of technology. From technical artefacts to sociotechnical system. Berlin: Springer; 2022.
14. van de Poel I. Embedding values in artificial intelligence (AI) systems. Mind Mach. 2020;30:385–409. https://doi.org/10.1007/s11023-020-09537-4.
15. Kroes PA, Meijers AWM, Franssen MPM, Houkes WN, Vermaas PE. The dual nature of technical artifacts. The Hague: Netherlands Organization for Scientific Research; 1999.
16. Floridi L, Sanders JW. On the morality of artificial agents. Mind Mach. 2004;14:349–79. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.
17. Johnson DG. Computer systems: moral entities but not moral agents. Ethics Inf Technol. 2006;8:195–204. https://doi.org/10.1007/s10676-006-9111-5.
18. Gunkel DJ. The machine question: critical perspectives on AI, robots, and ethics. Cambridge: The MIT Press; 2017. p. 68.
19. Vanderelst D, Winfield A. The dark side of ethical robots. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. p. 317–22. https://doi.org/10.1145/3278721.3278726
20. Latour B. We have never been modern. Cambridge, MA: Harvard University Press; 1993.
21. Latour B. On technical mediation: philosophy, sociology, genealogy. Common Knowl. 1994;3:29–64.
22. Latour B. 2002 Morality and technology: the end of the means. Theory Cult Soc. 2002;19:247–60.
23. Loh W, Loh J. Autonomy and responsibility in hybrid systems: the example of autonomous cars. In: Lin P, Abney K, Jenkins R, editors. Robot ethics 2.0: from autonomous cars to artificial intelligence. New York: Oxford University Press; 2017. p. 35–50. https://doi.org/10.1093/oso/9780190652951.003.0003.
24. Verbeek PP. What things do: philosophical reflections on technology, agency, and design. Penn State: Penn State University Press; 2005. p. 166.
25. Ihde D. Experimental phenomenology: an introduction. New York: State University of New York Press; 1977.
26. Ihde D. Technics and praxis. Dordrecht: Reidel; 1979.
27. Ihde D. Technology and the lifeworld: from garden to Earth. Bloomington and Indianapolis: Indiana University Press; 1990.
28. Verbeek PP. Obstetric ultrasound and the technological mediation of morality: a postphenomenological analysis. Hum Stud. 2008;31:11–26.
29. Verbeek PP. Moralizing technology: understanding and designing the morality of things. Chicago and London: University of Chicago Press; 2011. p. 108.
30. Ganguli D, Askell A, et al. The capacity for moral self-correction in large language models. arXiv:2302.07459v2 [cs.CL] 18 Feb 2023.
31. Coeckelbergh M. Introduction to philosophy of technology. New York: Oxford University Press; 2020. p. 67.
32. Redaelli R. Composite intentionality and responsibility for an ethics of artificial intelligence. Scenari. 2022;17:139–56.