Science and Engineering Ethics (2020) 26:2401–2410
https://doi.org/10.1007/s11948-020-00237-x

COMMENTARY

Expanding Nallur's Landscape of Machine Implemented Ethics

William A. Bauer[1]

Published online: 6 July 2020

Introduction

What ethical principles should autonomous machines follow? How do we implement these principles, and how do we evaluate these implementations? These are some of the critical questions Vivek Nallur asks in his essay "Landscape of Machine Implemented Ethics (2020)." He provides a broad, insightful survey of answers to these questions, especially focused on the implementation question. In this commentary, I will first critically summarize the main themes and conclusions of Nallur's essay and then expand upon the landscape that Nallur presents by suggesting additional approaches to machine ethics. The approaches I discuss reflect normative ethical theories and need not be applied only to machines, although machine ethics is the focus here. The overall goal is to open up further questions and research possibilities as society searches for the best approach to machine ethics.

Critical Summary of the Landscape

As Nallur conceives them, autonomous machines have the decision-making power to achieve goals by acting in the world. They are artificial agents. Because these machines impact humans, he reasonably asserts that they "need to be imbued with a sense of ethics that reflect the social milieu they operate in and make decisions that are ethically acceptable to society." The emphasis on society and the social milieu indicates Nallur's rather descriptivist approach to machine ethics; the idea is to understand what we actually want or what values are actually represented in society, then implement those in machines. We need machine ethics, he observes, because humans cannot be there to supervise or make every decision that machines need to make, and in fact, could slow processes down considerably when time is of the essence (e.g., in emergency situations).

✉ William A. Bauer
wabauer@ncsu.edu; wandbauer@gmail.com

[1] Department of Philosophy and Religious Studies, North Carolina State University, Raleigh, USA

First, we should clarify the nature and scope of the questions Nallur raises. This provides an opportunity to reveal two broad types of questions asked in the field of machine ethics and the ethics of artificial intelligence (AI), which will help clarify Nallur's project.

There are two types of questions within machine ethics. The first concerns small-scale interactions. Where there are relatively small groups of interacting agents (at least two), which might involve human agents, we want to specify which norms artificial moral agents should follow in their interactions.[1] These questions about small-scale interactions seem to be Nallur's primary focus, but beyond these, there are ethical questions about the large-scale, sociotechnical systems in which machines operate. These are large-scale issues concerning the democratic governance of sociotechnical systems. For example, there are investigations about how to put "society-in-the-loop" (Rawhan 2017) (modeled on the idea of a human-in-the-loop), there are questions about governance norms and "ethics in the large" (Chopra and Singh 2018), and there are worries about the socioeconomic implications of an increasingly AI workforce (e.g., see Dubljević and Bauer 2020). In response to these kinds of concerns, Crawford and Calo (2016, p. 313) advocate what they call "social-systems analysis," an approach to studying AI's impact on humans that "thinks through all the possible effects of AI systems on all parties" using a variety of humanities and social sciences disciplines. This kind of approach, I suggest, is needed to address the ethics of sociotechnical systems.

Although these domains of machine ethics—small-scale versus large-scale—may not be entirely separable, they do raise distinct kinds of questions. For example, a robot designed to assist the elderly will be capable of various actions regarding those under its care and we would like to ensure that its individual decisions are ethically justified, whereas an automated financial-trading system will be integrated with economic and other systems and thus raise larger-scale normative questions in the socioeconomic realm.

Machine actions at both scales should undergo ethical assessment and be subject to the most appropriate and rationally justified normative frameworks that we can develop. This is important for several practical reasons, among them to ensure good outcomes in our interactions with machines, which will, in turn, generate public trust in machines,[2] a desideratum that Nallur recognizes in his concluding section.

Nallur analyzes attempts at answering the smaller-scale questions concerning actions at the scale of individual machines.[3] In doing so, he makes a distinction

[1] These small-scale concerns include the ethics of autonomous vehicles, weapons systems, and medical assistants, as well as other similar examples. But there are also questions about the impact of AI and machines on various aspects of our personal and social lives, e.g., the impact of AI assistants (personal assistants such as Google Assistant, Alexa, and Siri) on autonomy (e.g., see Danaher 2018; Bauer and Dubljević 2019). Additionally, we can imagine ethical issues concerning the effects of AI technology on animals (Gary Comstock suggested this to me). For instance: Is it acceptable to replace a living pet with a robot pet, which might reduce the number of living pets in homes? Should we have AI devices monitor our pets while we are away?

[2] See Ferrario et al. (2019) for discussion of trust in AI.

[3] This is not to say that elements of Nallur's discussion do not have relevance to the large-scale issues.

between types of ethics and techniques used for implementing ethics. Regarding types of ethics, he introduces both consequentialist and deontological approaches. His example of a consequentialist ethical framework (specifically within machine ethics) is Asimov's three laws of robotics (Asimov 1950/2004). However, I think this should be classified as a deontological framework. The three laws are duties—e.g., "One: a robot may not injure a human being or, through inaction, allow a human being to come to harm."[4] This does not seem to have the flavor of a consequentialist demand, such as maximizing human happiness or maximizing the welfare of all sentient creatures.

In describing numerous implementation techniques—including the derivation of general rules from examples, using a rule-based governor, constraint-satisfaction techniques, formal mathematical verification, reinforcement learning, and hybrid ethical reasoning agents—Nallur appears to imply further distinctions that arguably reveal different types of ethics. For example, we can distinguish between rule-based versus act-based versions of utilitarianism (I will discuss utilitarianism more below). Something like this is implied, I think, in his discussion of rule-based governors and hybrid ethical reasoning agents. Moreover, the line between ethical theory and implementation technique is not always separate, for certain kinds of theories are best suited for certain kinds of implementations. For these reasons, I think that there is ample logical space to expand the landscape, as I will do in the next section.

Nallur concludes that none of the implementation attempts he discusses are robust enough, emphasizing that "[…] the same machine must be able to handle different contexts." In order to achieve domain robustness, I suggest that we need to require that machines implement both domain-specific and general ethical principles. Specific domains have their own moral desiderata, and a machine operating in a specific domain should be able to meet that domain's demands. But some machines will need to operate in multiple different domains, hence the need for general ethical principles for machines. (At the same time, a given machine is not always—rarely, in fact—detached from larger contextual considerations that apply across all or nearly all domains; hence there is a need for the kind of large-scale ethics previously mentioned.)

It is a truism that we should aim to build the most moral machines, and thus aim for the 'best' moral theory or principles. Given the array of options before us, perhaps a shotgun approach is useful. Nallur recommends competition amongst machines with different standards of ethics. May the most moral machine win! Build lots of different moral (or purportedly moral) machines, built on different theories or frameworks, then put them in the same environment to complete and see which moral approaches win out. Winning, of course, is going to depend on our judgments of success—which of two machines respects rights more consistently, saves the most

[4] Asimov first described the three laws in his classic short story "Runaround." Interestingly, in the story the rules are portrayed as creating potentialities for (or against) the types of actions they specify, and these potentialities can come into conflict. But they are pretty clearly deontological rules, specifying rather strict requirements. In sum, the three laws (in order of priority) are: do not harm humans, obey humans, and protect your own existence. For discussion of the three laws, see Thornton et al. (2017, p. 1431) and Wallach (2008).

lives, matches expert intuition, matches public opinion? This shotgun approach is all to the good, but we can still add further approaches to the competitive mix. The approaches I discuss below have the virtue of potentially being operable across multiple domains.

Expansion of the Landscape

How do we get artificial agents to recognize and respond appropriately in the space of moral problems? Typically, humans are capable of solving moral problems in a wide array of situations. To be sure, we make moral mistakes, and even when we think we have done what is morally right, frequently others will disagree. However, we want machines to be at least as good as we are. We need abstract theories or principles or methods that capture what it is, upon careful analysis, we think we *should* do and how we decide (either implicitly or explicitly) upon courses of action, so that in turn we can program these values and decision procedures into machines.

There are two broad approaches to implementing ethics: a top-down approach, where rules or general principles are imposed on the machine, and a bottom-up, 'figure out ethics as you go' approach. Both approaches are present in various forms in Nallur's discussion.[5] However, there are additional moral-theoretic frameworks, beyond those he discusses, available in the landscape of machine ethics. A complete and comprehensive mapping of all possibilities is not sought here; rather, expansion of the landscape is my goal.

Virtuous Artificial Moral Agents

Howard and Muntean (2016, 2017) advance a sophisticated model of a virtuous artificial moral agent. This is philosophically grounded in virtue theory as developed by Aristotle in *Nichomachean Ethics*. Their model emphasizes learning from moral examples to develop dispositional states, such that the agent will become more and more disposed to exhibit virtuous behavior. Their approach is intended only for artificial moral agents to simulate (not replicate) human virtue. The virtuous artificial moral agent is analogized to several aspects of human moral learning, most importantly that acting virtuously is a practical skill of humans that can be simulated in artificial agents through machine learning techniques (Howard and Muntean 2017, pp. 136–138). Moral cognition, in effect, can be modeled by machine learning, and the patterns that artificial agents can detect include patterns of virtuous behavior.

[5] See Wallach and Allen (2009) for discussion of top-down and bottom-up approaches, plus ways to hybridize them. The top-down/bottom-up approaches could be seen, I suggest, as mirroring Dennett's description of intelligent design theory as a 'skyhook' versus Darwinian theory as a 'crane': the former is a mind-first approach, not explainable by reference to prior existing material elements or forces, whereas the latter is a matter-first approach, explainable solely by references to material elements (Dennett 1995, p. 76). On the analogy, from the point of view of the machine, a top-down machine ethic would be forced on the machine (as a skyhook comes out of the blue) or developed bottom-up 'organically' through the machine's interactions (as a crane is built out of existing materials).

Howard and Muntean (2016, 2017), like many other scholars in the area of machine ethics, assume that moral functionalism will be good enough to create moral machines. We do not need to replicate the inner qualitative states of phenomenology of moral decision making, only the functional equivalent of these, to create moral machines.[6]

Note that rules or principles are not of primary importance to the virtuous artificial moral agent. Although this approach is "independent from pre-imposed, exceptionless rules," it does allow them while minimizing their importance (Howard and Muntean 2016, p. 221). The virtue-theoretic approach is one that deserves further consideration, but in contrast to it, there are approaches that emphasize the primacy of rules. Although Nallur discusses rule-based approaches, I would like to explore one rule-based approach that Nallur does not address.

Two-Level Utilitarian Artificial Moral Agents

Two-level utilitarianism is a nuanced utilitarian theory proposed by Hare (1983) that envisions two levels of thinking in our moral judgments.[7] The first level consists of intuitive, workaday rules for actions across a variety of domains; the principle of utility ultimately justifies these rules. The second level consists of critically applying the principle of utility to solve moral dilemmas to which no first-level rules apply. The principle of utility—that an action is morally right if and only if it maximizes happiness—ultimately reigns on the two-level approach. The rules are, at least on one plausible interpretation of the two-level approach, time-savers; if we had sufficient time and complete information about all the possible actions in response to some dilemma (and we were unbiased), the critical thinking level should be enacted to ensure that happiness is maximized.

Lucas and Comstock (2015) propose a satisficing hedonistic act utilitarian approach to machine ethics.[8] As they affirm (2015, p. 82), their approach is consistent with two-level utilitarianism. They accept the usefulness of the two-level approach for machine ethics and contrast their utilitarian proposal with the deontological, prima facie duties approach to ethics (Ross 1930) as exemplified in "EthEl," an AI medical assistant conceptualized by Anderson and Anderson (2008).

[6] Moral functionalism requires that "moral agents develop the disposition of replicating the semantics of a moral community by learning its folk morality"—the virtuous traits they develop are dispositions "nourished through a process of learning" (Howard and Muntean 2017, p. 135).

[7] See Varner (2012) for additional discussion and defense of two-level utilitarianism.

[8] Since their view is a *satisficing* one, utilitarian decisions should generate "a level of utility that leads to overall gains in happiness for some without costing anyone unhappiness" (Lucas and Comstock 2015, p. 81). Also, their model may not be fully implementable given current technology (2015, p. 92).

In Bauer (2018), I propose a two-level utilitarian approach to machine ethics on different grounds.[9] I develop two-level utilitarianism as a major competitor to the virtuous artificial moral agent advanced by Howard and Muntean (2016, 2017) (discussed above), arguing that the two-level approach satisfies some of the main desiderata that a virtue-theoretic approach demands. Importantly, the rules that an artificial moral agent might contain have a semblance to virtue. According to virtue theory, virtues are dispositional states which have a rule-like structure, for they can be operationalized as follows: 'given stimulus S, in conditions C, take action A.' Hooker (2000, p. 90) argues, and I agree, that rules can become sufficiently ingrained in one's character such that "to accept a code of rules is just to have *a moral conscience of a certain shape*."[10] By analogy, I suggest that machines can develop 'character' with a sufficiently robust set of rules (or dispositional states).

There is an important difference in the two-level utilitarian approaches separately discussed by Lucas and Comstock (2015) and Bauer (2018). Lucas and Comstock (2015, p. 88) suggest that an artificial moral machine use level two first, then default to level one if insufficient time is available to determine a course of action.[11] By contrast, I emphasize that level one should, at least for most practical matters, come first; the rules therein can consist of initial rules for various specific domains, or more general rules. Then, only when the system lacks sufficient rules to determine a course of action, and if sufficient information and time are available to make a definitive calculation, level two (the utility calculator) kicks into action. The utility calculator would determine the best course of action, and in turn, this would allow the artificial moral agent to build up further rules for future action.

Perhaps the two-level utilitarian approach could be implemented by the kind of rule-based governor system that Nallur discusses, which incorporates a "constraint module that either allows or disallows actions to be taken by a particular system, based on whether the action would break pre-existing rules." The benefits of

[9] I stated in Bauer (2018) that as far as I knew no one had applied two-level utilitarianism to machine ethics. However, Lucas and Comstock (2015) had done so. Nonetheless, there are important differences in the way I developed it. First, mine was a general machine ethics proposal, whereas Lucas and Comstock seemed more focused on medical assistant robots (although the approach has more general applications); second, I was concerned to compare and contrast my proposal with the virtue ethics approach, whereas Lucas and Comstock did so with the prima facie duties approach to ethics; third, level two seems to have priority in their version, whereas level one has priority in my version (more on this in the text below).

[10] Hooker (2000, p. 89), however, does not accept two-level utilitarianism or a strict act-utilitarianism; he advocates only for rule-utilitarianism (for discussion of major differences between act, rule, and two-level utilitarianism, see Bauer 2018). The problem, he argues, with adding an act-utility calculator on top of a system of rules is that, when conflicts between rules need to be resolved by the act-utility calculator, people would tend to lose confidence in the rules. I think, however, that confidence could be further inspired by having the second level, the utility calculator, in place in order to definitively resolve conflicts.

[11] Consistent with the two-level approach, Lucas and Comstock (2015, p. 86) also recognize that following Ross' prima facie duties is advisable at first, but as information grows, the machine should follow hedonistic act utilitarianism. However, it seems that in their ideal ethical machine the utility calculator would get priority, and only resort to level one rules if there is insufficient time to calculate the best possible course of action (2015, p. 88).

two-level utilitarianism are that it bridges deontological and consequentialist ethical dimensions and can simulate virtuous behavior. When new rules are learned as the machine uses its utility calculator, the rules, which have the feel of duties, are internalized in a way akin to acquiring the dispositional states essential to virtue. So, in a sense, the two-level utilitarian approach represents a kind of pluralistic approach. However, ultimately, the principle of utility reigns.[12]

What would a genuine pluralistic approach look like? Dubljević and Racine (2014) propose a pluralistic model of moral decision making that incorporates the 'big three' moral traditions: virtue theory, deontology, and utilitarianism. It is called the ADC model (for agent, deed, consequence). This model has been empirically explored (Dubljević et al. 2018) and could be developed for machine ethics; Dubljević (2020) goes some way towards that in his contribution to this Topical Collection. Additionally, there are proposals for specific domains that could potentially be generalized and assessed from an implementation point of view. For example, Leben (2017) uses Rawls' theory of justice as the basis for an approach to autonomous vehicle ethical decision-making.

Rawlsian Artificial Moral Agents

Working within the social contract theory tradition, Rawls (1971) developed a theory of 'justice as fairness'. The theory concerns large-scale socioeconomic justice. It applies to situations where the question of how we should distribute harms and benefits, or obligations and rights, across society is at stake. It is like utilitarianism in having direct implications for large-scale social issues, but strongly disagrees with the utilitarian demand to maximize the good (be it happiness, well-being, or what have you). Rawls asks us to follow the maximin principle: given that inegalitarian outcomes are permitted, it is imperative to institute a distribution of benefits such that those in the least privileged (or most harmed) group are in fact benefitted as much as possible. Such an approach, for Rawls, is justified by his 'original position' thought experiment: what distribution of goods (both immaterial and material) would you choose if you were selfish but did not know your personal details (class, gender, wealth level, etc.)? That is, how would you construct society if you were behind a 'veil of ignorance' about your personal characteristics—if you did not know who you were?

As Leben (2017, p. 109) notes, for Rawls all participants will agree on rules from the original position "since they are all effectively the same player." Leben (2017) breaks new ground by employing the maximin principle, a core aspect of Rawlsian justice, to the ethics of autonomous vehicles. From behind the veil of ignorance, you should remain agnostic about which actions better serve your (non-veiled) interests because, for all you know, you could be riding in the autonomous vehicle or be a pedestrian (2017, p. 112). As such, Leben develops an algorithm for Rawlsian decision making that requires the implementation of a maximin procedure. It picks

[12] For further discussion of utilitarian machine ethics, see Grau (2006).

the highest payoff from the set of lowest payoffs, and this keeps running until there are no payoffs available. The distribution of harms/benefits in AV crashes is just one domain,[13] but the maximin approach to decision-making is relevant to other domains as well (healthcare decisions made by medical robots is another example). The best thing about a Rawlsian algorithmic approach is "its respect for persons as equals, and its unwillingness to sacrifice the interests of one person for the interests of others" (Leben 2017, p. 114). Rawls' theory is a political theory that incorporates insights from moral theory—e.g., Rawls incorporates Kant's conception of respect for persons as distinct, self-legislating beings[14] and has major potential implications for a variety of moral issues (as Leben brings out). But more generally, what are the connections between political theory and artificial moral agents?

Political Theory and Artificial Moral Agents

Himmelreich (2020) argues for the need for political theory in thinking about the ethical implications of AI and technology. In particular, we should explore all questions of small-scale machine ethics against our sociopolitical structures (the large-scale concerns I discussed earlier). Thus, we need political theory, not just moral theory, if we are to implement the best machine ethics.

Ethical considerations cannot be separated from the question of justice in sociopolitical institutions. If we value autonomy, for instance, we need public policy that permits choice in how individual machines behave, e.g., allowing users of autonomous vehicles to personally set some of the driving parameters (Himmelreich 2020, p. 35). According to Himmelreich, policy and institutional questions are just as relevant, if not more relevant, to AI's impact—and not just to the larger questions but also to the smaller-scale questions. We should keep in mind that moral theory and political theory (and morality and politics) are not entirely separate subjects. For instance, utilitarianism is often portrayed as an ethical theory in textbook discussions, but it is also a political theory—e.g., it is the main view that Rawls (1971) criticizes in advancing his theory of 'justice as fairness.'

In particular, political philosophy can offer three concepts identified by Himmelreich (2020, p. 35) that are relevant to machine ethics: reasonable pluralism (humans justifiably disagree about complex issues, given our disparate life experiences), autonomy (since individuals often know what is best for themselves, individual choice should in some cases be given higher weight than socially preordained values), and legitimate authority (who decides what is regulated, and how they do so, is relevant to a given policy's public and moral acceptability). With these concepts at the forefront of our efforts to implement machine ethics, we are more likely to create a value framework acceptable to a diverse society.

[13] Leben provides detailed comparisons of three AV scenarios that he envisions.

[14] By contrast, utilitarianism is often criticized as allowing violations of individuals' rights.

Concluding Remarks

Nallur's landscape of machine ethics can be expanded in two ways. First, it can broadened by including large-scale ethical issues as well as small-scale ethical issues. These are actually interconnected, as I have shown at various points in my discussion. Second, it can be deepened (made even stronger) by including additional moral-theoretic grounded machine ethics proposals, such as virtuous and two-level utilitarian artificial moral agents.

Nallur devotes considerable space to discussing methods for testing ethics, but I have not focused on that in my commentary. As he asserts, given the emphasis on validation in computer science and robotics research, "there has been more discussion about how implementations were to be evaluated, rather than on which ethical theory is a better candidate for implementation." I have tried to remedy that a bit by offering up further theories for testing. How their implementation is to be tested needs further consideration. Attempting to resolve the kinds of dilemmas identified by Nallur (the trolley problem, cake or death, the burning room, and so on) is at least part of a feasible testing approach. Successful, culturally acceptable integration of autonomous machines into society will be another test.

Acknowledgements Thanks to Gary Comstock for helpful suggestions concerning several specific points, and especially concerning two-level utilitarianism. Thanks to Veljko Dubljević for feedback about the ADC model of moral judgment.

References

Anderson, M., & Anderson, S. L. (2008). EthEl: Toward a principled ethical eldercare robot. In *Eldercare: New Solutions to Old Problems*. Proceedings of AAAI Fall Symposium. Washington, D.C. https://www.aaai.org/Library/Symposia/Fall/fs08-02.php.

Aristotle. 350 BCE. *Nicomachean Ethics*. W. D. Ross (Trans.). *The Internet Classics Archive*. https://classics.mit.edu/Aristotle/nicomachaen.html.

Asimov, I. (1950/2004). *I, Robot*. New York: Random House.

Bauer, W. (2018). Virtuous vs. utilitarian artificial moral agents. *AI & Society, 35*, 263–271. https://doi.org/10.1007/s00146-018-0871-3.

Bauer, W., & Dubljević, V. (2019). AI assistants and the paradox of internal automaticity. *Neuroethics*. https://doi.org/10.1007/s12152-019-09423-6.

Chopra, A. K., & Singh, M. P. (2018). Sociotechnical systems and ethics in the large. In *AIES '18, Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society* (pp. 48–53). https://doi.org/10.1007/s12152-019-09423-6.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*, 311–313. https://doi.org/10.1038/538311a.

Danaher, J. (2018). Toward an ethics of AI assistants: An initial framework. *Journal of Philosophy and Technology, 31*(4), 629–653. https://doi.org/10.1007/s13347-018-0317-3.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meaning of life*. New York: Simon & Schuster.

Dubljević, V. (2020). Toward implementing the ADC model of moral judgment in autonomous vehicles. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00242-0.

Dubljević, V., & Bauer, W. (2020). Autonomous vehicles and the basic structure of society. In R. Jenkins, D. Černý, & T. Hříbek (Eds.), *Autonomous vehicles ethics: Beyond the trolley problem*. Oxford: Oxford University Press.

Dubljević, V., & Racine, E. (2014). The ADC of moral judgment: Opening the black box of moral intuitions with heuristics about agents, deeds and consequences. *AJOB Neuroscience, 5*(4), 3–20.

Dubljević, V., Sattler, S., & Racine, E. (2018). Correction: Deciphering moral intuition: How agents, deeds and consequences influence moral judgment. *PLoS ONE, 13*(10), e0206750. https://doi.org/10.1371/journal.pone.0206750.

Ferrario, A., Loi, M., & Viganò, E. (2019). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00378-3.

Grau, C. (2006). There is No "I" in "Robot": Robots and utilitarianism. *IEEE Intelligent Systems, 21*(4), 52–55.

Hare, R. M. (1983). *Moral thinking: Its levels, method, and point*. Oxford: Oxford University Press.

Himmelreich, J. (2020). Ethics of technology needs more political philosophy. *Communications of the ACM, 63*(1), 33–35.

Hooker, B. (2000). *Ideal code, real world*. Oxford: Oxford University Press.

Howard, D., & Muntean, I. (2016). A minimalist model of the artificial autonomous moral agent (AAMA). *Association for the Advancement of Artificial Intelligence.*

Howard, D., & Muntean, I. (2017). Artificial moral cognition: Moral functionalism and autonomous moral agency. In T. M. Powers (Ed.), *Philosophy and computing, philosophical studies series* (Vol. 128, pp. 121–160). Berlin: Springer.

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology, 19*, 107–115.

Lucas, J., & Comstock, G. (2015). Do machines have prima facie duties? In S. P. van Rysewyk & M. Pontier (Eds.), *Machine medical ethics* (pp. 79–92). Berlin: Springer. https://doi.org/10.1007/978-3-319-08108-3_6.

Nallur, V. (2020). Landscape of Machine Implemented Ethics. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00236-y.

Rahwan, I. (2017). Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology, 20*, 5–14.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Belknap.

Ross, W. D. (1930). *The right and the Good*. Oxford: Oxford University Press.

Thornton, S. M., Pan, S., Erlien, S. M., & Gerdes, J. C. (2017). Incorporating ethical considerations in automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems, 18*(6), 1429–1439.

Varner, G. (2012). *Personhood, ethics, and animal cognition: Situating animals in Hare's two level utilitarianism*. Oxford: Oxford University Press.

Wallach, W. (2008). Implementing moral decision making faculties in computers and robots. *AI & Society, 22*, 463–475.

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

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