To appear in Markus Dubber, Frank Pasquale, and Sunit Das (Eds.), *Oxford Handbook of Ethics of AI*, Oxford University Press, 2020.

Abstract

This chapter explores a human-centered approach to AI and robot ethics. It demonstrates how a human-centered approach can resolve some problems in AI and robot ethics that arise from the fact that AI systems and robots have cognitive states, and yet have no welfare, and are not responsible. In particular, the approach allows that violence toward robots can be wrong even if robots cannot be harmed. More importantly, the approach encourages people to shift away from designing robots as if they were human ethical deliberators. Ultimately, the cognitive states of AI systems and robots may have a role to play in the proper ethical analysis of situations involving them, even if it is not by virtue of conferring welfare or responsibilities on those systems or robots.

Keywords

human-centered approach, AI ethics, robot ethics, AI systems, cognitive states, robots, ethical analysis, mens rea

To appear in Markus Dubber, Frank Pasquale, and Sunit Das (Eds.), *Oxford Handbook of Ethics of AI*, Oxford University Press, 2020.

A Human-Centered Approach to AI Ethics:

A Perspective from Cognitive Science

Ron Chrisley

The increasing role of artificial intelligence (AI) and machine learning technology in our lives has raised an enormous number and variety of ethical challenges, as can be seen in the diverse topics covered in this volume. In addition, there are the ethical challenges yet to come, ones that we cannot currently anticipate. We can try to respond to this vast array of challenges individually, in an ad hoc manner, but in the long run, a more principled, structured response is likely to be of more guidance. In this chapter I propose responses to some particular questions concerning the ethics of AI, responses that share a unifying perspective: a human-centered approach. The hope is that, beyond offering solutions to the particular problems considered here, these responses can be of more general interest by illuminating enough of their shared, human-centered perspective to facilitate like-minded responses to any number of current and future ethical challenges involving AI.

More will be said about what the human-centered approach to AI/robot ethics amounts to, but an important consequence of it, and the central claim of this chapter, is this: when making ethical judgments in this area, we should resist the temptation to see robots as ethical agents or patients. For the foreseeable future, more ethical hazard follows from seeing humans and robots as ethically analogous than follows from seeing them as ethically distinct kinds. Much of what I say in what follows is meant to support this claim, to identify some instances of current practice that fail to heed the warnings of the claim and to suggest ways of avoiding the anthropomorphic error the claim identifies, while still minimizing the likelihood of certain ethically adverse outcomes involving robots and AI in general.

This central claim can seem at odds with an otherwise attractive naturalism about ethics, mind, and what it is to be human. My adoption of a human-centered approach to the ethics of AI arises out of my lifelong interest in cognitive science. Cognitive science is the interdisciplinary search for an understanding of how mentality in general (not just cognition) can be part of the natural world and the use of that understanding to provide explanations of mental phenomena and the behavior of systems with minds. One might think that this naturalism (particularly in the mechanistic, functionalist, physicalist form that many traditional cognitive scientists embrace, even if only implicitly) encourages us to see ourselves as glorified robots, a rough equation that would either support the extension of the concepts of ethical agent and patient to suitably programmed robots and AI systems, or encourage ethical nihilism for both humans and robots. Contrary to this, I believe that seeing humans as part of the natural world does not undermine our understanding of what makes humans ethically different from robots (or nonhuman animals); rather, it gives that understanding scientific plausibility and conceptual clarity. It is only by properly considering our place in the natural world that we can see the true, nondualist, reasons why it is correct to see us, but not robots (at least for the forseeable future), as ethical beings. Nevertheless, the theories and methods of cognitive science will largely remain in the background of this chapter, with the focus instead being on the human-centered approach they support.

Putting Robots in Their Place

Just what do I mean by a human-centered approach? We'll be better equipped to answer that question in full after we have a few instances of it from which to generalize, but a few things can be said at the outset to give an initial idea of what the approach is—and what it is not.

The human-centered approach to AI ethics I am advocating here has two key aspects:

1. An emphasis on human welfare.

2. An emphasis on human responsibility.

The first aspect is in contrast with approaches to AI ethics that take seriously ethical obligations concerning the purported welfare of artificial agents. Such approaches focus on questions such as:

- Can robots feel pain?
- Can they suffer?
- If so, what are our obligations, if any, for reducing robot pain and suffering at the expense of increasing human pain and suffering? At the expense of increasing animal pain and suffering?

Similarly, the second aspect of the human-centered approach is in contrast with approaches that focus on questions such as:

- Should we punish robots that are responsible for crimes?
- Should we grant citizenship, workers' rights, "human" rights to certain machines?
- Should robots be allowed to own property?

The human-centered approach doesn't just answer questions like these in the negative; it dismisses them as impertinent, or worse: as presupposing a view that is so wrong-headed that it risks both distracting us from many of the real ethical issues, and misdiagnosing those few real issues we do manage to address.[1] The human-centered approach starts with the following *Deflationary View* about machine ethics:

[1] For a human-centered AI ethics from a substantially different perspective, see, e.g., Bryson, Joanna. 2018. "Patiency is not a virtue: the design of intelligent systems and systems of ethics". *Ethics and Information Technology*. 20 (1): 15-26.

Deflationary View: No robot or AI system currently in existence could be ethically responsible or be the kind of thing toward which we have ethical obligations.[2]

Some might be inclined to stop reading at this point, believing I have just dismissed, without argument, most of what is of interest in AI ethics. Fair enough; the previously posed questions are alluring and excite our imaginations, so interest in them is understandable. And attempting to answer such questions can be a good way to explore the features and limits of the concepts involved in stating them. But if it is worthwhile to consider ethical issues that arise in futuristic thought experiments involving AI and robots, it is all the more worthwhile, even pressing, to consider the ethical issues confronting us now, in a way that is not unduly distorted by consideration of the counterfactual, futuristic, robot-as-ethical-agent-or-patient cases.

I do not wish to be confused for an AI pessimist, so let me make one thing explicitly clear: the Deflationary View applies to AI systems/robots currently in existence (or in the foreseeable future). In taking the Deflationary View, the human-centered approach I am advocating does not thereby assume that only humans (or beings biologically related to humans: animals) could ever be ethically responsible agents or deserving of our ethical concern. For example, I am not advocating the Deflationary View because I believe there is some fundamental inability for artifacts (or nonbiological systems, whatever their provenance) to have minds, to experience emotion, to be conscious; on the contrary. My point is that while in principle, there might someday be robots or AI systems that are ethically responsible or are the kinds of things toward which we have ethical obligations, in fact there are not nor are there likely to be in the foreseeable future.[3] Unlike an AI ethics that addresses

[2] Perhaps unsatisfyingly, I do not argue for this claim here. One reason for thinking that current AI systems are not moral agents is that they lack the capacity for *judgment* (in a specific, almost technical sense of that word; see Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence. Reckoning and Judgment*. Cambridge: The MIT Press. 124-127.).

[3] Thus while others may be correct in their accounts of what conditions would have to be met by an AI system or robot for it to enjoy ethical status (e.g., Sullins, John. 2006."When Is a Robot a Moral Agent?". *International*

the previously posed questions, a human-centered AI ethics is urgently needed, now. And since it seems likely that we will continue to use AI systems and robots that are not responsible nor to which we have any responsibility, even beyond the eventual advent of AI systems with their own ethical status, a human-centered AI ethics will continue to be indispensable even if a more substantive AI ethics, based on the obligations of and toward AI systems and robots, becomes necessary.

The human-centered approach to AI ethics I am advocating is deflationary in another, related aspect. Some hold that current AI and machine learning is an ethical game changer, which requires a radical break in our ethical thinking in order to accommodate artificial agents that are responsible for their actions and/or to which we bear some responsibility. The human-centered approach being offered here is conceptually conservative, urging us to try to use precedent, past wisdom, and conventional metaphysics as much as possible when trying to resolve ethical issues involving current and near-future AI technology. On such an approach, robots and AI systems, despite any autonomy, learning or decision-making capabilities they may have, are best treated, in our ethical deliberations, in a manner continuous with how we deal with other technologies: as nonpersonal boundary conditions potentially affecting the praise- or blame-worthiness of the people involved—not as candidates for such praise or blame, nor as personal subjects whose harm or benefit can figure, in the special way personal well-being does, into the ethical evaluation of human action. On the other hand, we cannot afford to be complacent. These new, highly adaptive and flexible technologies are unlike any before and require new ethical concepts and tools. But the new concepts and tools we need should not be developed by diagnosing our situation in terms of the arrival on the ethical scene of a new source, or target, of ethical responsibility.

Review of Information Ethics. 6: 24-29.), it is my view that these conditions will remain unmet for the foreseeable future.

If the questions listed previously are not the right or relevant ones, what are? Here are a few:

- To what extent should damaging, stealing, or destroying an adaptive information system "implant" that a person has trained over several years, and on which that person relies to function in everyday life, count as harm to that person, over and above the usual harm associated with property loss?
- When an autonomous robot takes action that results in harm or loss, how should the responsibility for that harm be distributed across the various people and organizations involved, such as:
    - The robot operator(s),
    - Bystanders,
    - The robot trainer(s),
    - The robot programmer(s),
    - The robot manufacturer(s),
    - The robot retailer, and
    - The governmental body that licensed robot operation in that context?
- In what ways can the use of certain kinds of augmenting AI technology better enable us to perform ethically? What AI technologies might instead compromise our ethical competence?

These questions are good indications of how to apply the human-centered perspective; they focus exclusively on the welfare and responsibility of the only ethical agents on the scene: humans.[4] But in some situations it can be tricky to see how to achieve this focus properly.

[4] This isn't quite right: animals are also "on the scene", so the impact of AI and robots on them should also be taken into consideration.

The remainder of this chapter will look at two kinds of case, the better to flesh out the human-centered approach. Until now, I have been referring to my area of interest using the cumbersome phrase "AI systems or robots," which is fair enough, since the human-centered approach to AI ethics applies that broadly. But for the remainder of this chapter I will just use the phrase "robots" and focus especially on social robots (ones designed to interact with humans, as opposed to, say, industrial assembly-line robots). It is when social robots are on the scene, much more so than in cases involving disembodied AI, that the temptations of an inflationary ethics, and the concomitant need to keep hold of the insights of the human-centered view, are at their strongest. Thus, a focus on social robots will make it easier to see the points I wish to make (and will streamline the prose). But the insights I will thereby uncover apply, I believe, to the more inclusive class of AI systems and robots in general.

Implications of a Human-Centered Approach

Corresponding to the two aspects of the human-centered approach identified at the outset, welfare and responsibility, I will look at two nonobvious or counterintuitive implications of the approach.

Harming Robots

The issue of harming robots can be emotionally charged and divided: for an example one only has to look at David Harris Smith and Frauke Zeller's hitchBOT, the actions of the authors of its fate, and people's responses to that treatment. Another example is the forceful actions Boston Dynamics uses to demonstrate the robustness of their robots, and people's emotionally charged reactions upon seeing videos of these demonstrations.

One might think that proponents of the human-centered approach to AI ethics have their hands tied here: according to the Deflationary View, robots are not the kind of thing that can be harmed, and so the question is dismissed as being immaterial to today's pressing ethical concerns with AI.

But the issue cannot be dismissed so easily: so that talk of "harm" does not beg the question, let's make it clear that in this context we mean it to cover actions that are of a kind that, were they performed on humans or animals, would cause harm. In more neutral language: is it ethical to hit, disfigure, mutilate, and so on robots? Further, we are concerned here with ethical prohibitions, if any, over and above those having to do with damaging someone's property in general.

It may seem that the human-centered approach proponents are still bound here: since such actions would not cause harm, they are not prohibited.

But is it really the case that such actions cause no harm? The mere consideration of whether the robot's welfare is relevant, even if answered in the negative, has done its own harm: distracted us from proper consideration of the humans in the situation (the agent of the harm and any observers). Because even if mutilating a robot does not harm the robot (because the robot is not the kind of thing that can be harmed), such mutilation may in fact do harm to the humans involved—an emphasis that is at the heart of the human-centered approach. The idea here says something about why it is wrong to abuse robots that is very similar to what Immanuel Kant says about why it is wrong to abuse animals (but does not commit anyone to agreeing with Kant on why animals should not be abused). The idea is that even if robots cannot be harmed, they are, at least sometimes, "made in our image" to such an extent that willfully abusing them is at best grotesque, at worst unethical. Think of how we would consider it grossly inappropriate for someone to willfully and sadistically (i.e., not as part of performance art, or as an experiment, or as a political protest, etc.) dismember a doll

(as opposed to, say, a toy car) on stage in front of young children. The key feature here is the doll's sharing, to some degree, the human form. Robots, to be sure, can share this visual form as much as any doll. But beyond that, they can share the human form, in a higher, more abstract sense, to a much greater degree: witness their ability to respond to questions with linguistic sounds, to acquire information from their environment and act conditionally upon it, to learn, decide, remember, prefer, assist, make emotional displays, and so forth. So, it could be argued, acts of violence upon these robots conceivably cross into the unethical because they brutalize the agent (and perhaps those witnessing the act). Accounts differ as to why harming something with the human form is wrong, with the familiar consequentialist (the normalization of violence to the human form makes violence to actual humans more likely) and deontological (it's just wrong to do harm to the human form) variants. As should be clear, I am not attempting to make a strong case for this view here; I am only pointing to the view as an existence proof that one can take the human-centered approach to AI ethics and still hold that it is unethical to abuse (some) current robots, in some situations.

But I also want to highlight another point that arises out of this discussion: although it was only the welfare and responsibility of humans that ultimately mattered in this case, the mind-like cognitive abilities of robots also played a crucial role. That is what made the issue one in the ethics of AI, rather than ethics in general. What's important to note is that the role those abilities played was not that of making the robots the kind of thing that could be harmed (or the kind of thing that could be responsible). Rather, the role it played, and what it is novel about such technology, involves the new and complex ways robots can impact on human welfare and human responsibility.[5]

[5] For more on the ethics of robot abuse, see, e.g., Whitby, B. 2008. "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents". *Interacting with Computers*. 20 (3): 326-333. See also Cappuccio, Massimiliano L., Anco Peeters, and William McDonald. 2019. "Sympathy for Dolores: Moral Consideration for Robots Based on Virtue and Recognition". *Philosophy & Technology*.

Robots as Extensions of Human Responsibility

Taking the human-centered approach to AI ethics can have practical consequences for robot design. This can be demonstrated by considering what such an approach has to say about one way of designing robots, which I call *logic-based ethical robot methodology*. You can think of this methodology as a direct descendant of the approach explored in Isaac Asimov's novels.

Logic-based ethical robot methodology:

- An ethical system is encoded (by humans) in logic;
- Robots are given these statements and an ability to reason logically with them;
- Robots consult these rules when generating their behavior (e.g., by disqualifying a proposed action if it is a consequence of their reasoning that the action is not ethically permissible).

Such robots are *explicit ethical agents* (sometimes called *explicit moral agents*) in James Moor's sense: "Explicit ethical agents are agents that can identify and process ethical information about a variety of situations and make sensitive determinations about what should be done. When ethical principles are in conflict, these robots can work out reasonable resolutions."[6]

The intended outcome of this methodology is not only avoidance of ethically adverse situations but also an ability to explain/justify robot behavior by appealing to the inferential trace that governed its generation.

6 James Moor, "Four Kinds of Ethical Robots," Philosophy Now 72 (March/April 2009): 12.

An example of work that employs the ethical robot methodology comes from Matthias Scheutz and Bertram Malle.[7] There, the authors consider ethical questions such as: should Rob, the elder-care robot, deliver pain medication even though it cannot consult a supervisor, as is usually required? The authors say:

"An interesting question is what a human health care provider might do in Rob's position . . . If R were to model human behavior, it would, in addition to ethical reasoning, need the capability for empathy as well as the ability to generate justifications (i.e., explanations of norm violations such as not contacting the supervisor). We will not focus on those aspects of moral competency in this paper. Rather, we will develop a general argument that, in order to avoid unnecessary harm to humans, autonomous artificial systems must have moral competence."[8]

In line with logical ethical robot methodology, Scheutz and Malle aim to give their robots said moral competence by giving them a set of logical axioms, some logical statements encoding the state of the world, and an ability to draw inferences from these:

1. ¬havePermission(R, administer(R, H, M)) ⟶ O[¬administer(R,H,M)] [obligation]
2. inPain(H) ⟶O[administer(R,H,M)] [obligation]
3. ¬havePermission(R, administer(R, H, M)) [fact]
4. inPain(H) [observation]
5. O[¬administer(R, H, M)] [1,3,MP]
6. O[administer(R, H, M)] [2,4,MP]
7. ¬◊(administer(R, H, M) ∧ ¬administer(R,H,M)) [modal logic]

[7] Matthias Scheutz and Bertram Malle, "Think and Do the Right Thing: A Plea for Morally Competent Autonomous Robots," *ETHICS '14: Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology* (IEEE: Chicago, 2014), 36–39. See also Malle B.F. 2016. "Integrating robot ethics and machine morality: the study and design of moral competence in robots". *Ethics and Information Technology*. 18 (4): 243-256.
[8] Ibid., 36.

Key: ¬ = negation, → =,material implication, ∧ = conjunction, O = it is obligatory that, ◊ = it is physically possible that, R = robot, H = human, M = medicine, MP = modus ponens. From (Scheutz and Malle 2014).

Lines 1 and 2 are the axioms, lines 3 and 4 encode facts about the world, and lines 5–7 follow deductively from the lines before them. This reasoning reveals a dilemma, in that it is both obligatory that the robot administer the medicine and that it not administer the medicine. The point here is not to consider the dilemma and its possible solutions; rather, this reasoning is only presented so that we can have to hand a concrete exemplar of the logic-based ethical robot methodology.

One might think that logical ethical robot methodology is in direct conflict with the human-centered approach. Robots can't have moral competence, one might say, because they have no responsibility. It makes no sense for them to reason about what is permitted or obligatory for them, because they have no obligations or permitted actions.

But is there some way to find a rapprochement between logical ethical robot methodology and the human-centered approach? One can reject the human-centered response, above, as heavy-handed, as misconstruing the meaning of the axioms Scheutz and Malle have provided. One need not read "O[¬administer(R, H, M)]" as encoding "it is obligatory, *for the robot*, that the robot administer the medicine to the human." Instead, one could read the statement as merely encoding the proposition that the state of affairs in which the robot administers the medicine to the human is ethically obligatory. This is a general statement of the ethical landscape, not tied to any particular agent. What is an obligation for one, is an obligation for all.

In the simple kinds of ethical systems and situations Scheutz and Malle are addressing, this picture may be adequate. But it cannot be adequate in general; we differ in our obligations and permissions, so reasoning in the abstract about what states are or are not

allowed to obtain will be of little to no use when deciding how to act ethically. One must in addition know where one is located in that web of obligations and permissions. The problem for the robot is that it is not located anywhere in that web. So that web can have no imperative force on its actions, even less so a force that could be inferred through logical reasoning.

The proponent of the logical ethical robot methodology could instead reply that it doesn't matter that the robot doesn't actually have the obligations it is reasoning about; all that matters is that the robot, by engaging in the kind of reasoning that would be correct were it a human, arrives at the ethically correct behavior (compare the distinction between genuine vs. functional ethical status[9]). To get the desired result, the robot need not *be* an ethical agent; it only needs to simulate one, to act *as if* it were an agent with obligations, and so on. That is what will get the right outcome. But given our differences in permissions and obligations, one has to ask: *which* ethical agent should the robot simulate?

To make the difficulty here clearer, consider: the inferences in the logical ethical robot methodology are not just used in the generation of behavior, but in the explanation/justification of it. But since (as we have been assuming all along), robots cannot be responsible, the justifications generated by the methodology should apply to the actions of humans, not robots. So "as if," robot-framed justifications will only be of use if they can help us construct actual, human-framed ones. But how are we to do this? As far as I can tell, the logical ethical robot methodology is silent on this issue. This failure to find a mapping from robot faux justifications to actual human justifications is itself a moral hazard, as it will lead to "moral murk." That is, everyone interacting with, writing about, training, making policy concerning, deploying, developing software for, designing, and so on will be invited to take

9 Steve Torrance and Ron Chrisley, "Modelling Consciousness-dependent Expertise in Machine Medical Moral Agents," in Machine Medical Ethics, ed. Simon Peter van Rysewyk and Matthijs Pontier (Berlin: Springer International Publishing), 295.

the attributions of responsibility to the robot at face value, given the lack of guidance on how to allocate that responsibility to the humans involved.

This is the heart of the matter, but, stated so abstractly, grasping its insight can be difficult. To see exactly how the logical ethical robot methodology can fail to properly allocate responsibility, and what a human-centered approach must do to remedy that deficiency, a specific example will be helpful. In particular, the important issues can be identified in a situation in which the epistemic state (mens rea) of the humans involved is a crucial component in evaluating the ethical status of their actions.

Consider an autonomous military robot R in a war zone with bridges A and B. The robot is under the command/control of human H. H can deploy R to patrol the region that contains bridges A and B. Among the actions that R can perform is the destruction of a given bridge. In this situation, it is in general an ethical good to destroy bridges, as it would protect innocents from attack—unless the bridge has a mini hospital with medical supplies on it, in which case the bridge should not be destroyed. Accordingly, R is designed such that if, even while out of contact with H, it acquires the information that it is very likely that there is no hospital on a bridge, it will destroy that bridge.

At the time of deployment, H believes that it is very likely that bridge A has no hospital on it, but that B does. So H deploys R. Soon after deployment, R loses contact with H and must rely on the reasoning given to it via the logical ethical robot methodology. On the way to bridge A, passing by bridge B, R acquires the information that it is very likely that there is no hospital on bridge B (by assessing B with its cameras, say). So it destroys the bridge. Unfortunately, and despite the information R received, there *was* a hospital on the bridge.

Should H be held responsible for the destruction of the hospital? The logical ethical robot methodology is silent on the issue: the only responsibilities it deals in are the "as if" responsibilities of either a robot which has none or an amorphous, unidentified human subject

of unknown identity who, had they been the one who had made the wrong call on destroying the bridge, might or might not be "let off," given that they acted in the best way with the best information at the time. But what do either of these have to do with the responsibilities of H (or the responsibilities of the designers of the algorithm that incorrectly assessed the status of the bridge)?

A more human-centered approach to this situation can be arrived at in one of two ways.

The first way is to keep the logical ethical robot methodology intact, but supplement it with a human-centered interpretative scheme. In the abstract, we have a situation in which H performs an action (deploying R) that results in an ethical disaster: the destruction of the hospital. We have a mitigating, mens rea–involving story, but that involves the epistemic state of R, not H, so as things stand, it cannot serve to reduce H's culpability.[10]

But perhaps things should not stand? What would it take for the information that R gleans while out of contact with H to mitigate H's culpability? Something like this: externally individuated epistemic states for subjects who are using autonomous epistemic technology, such as R. That is, for the purposes of determining H's culpability, H's epistemic state is to include the information gleaned by R, even while H and R are not in causal connection with each other. This allows us to arrive at what many would consider the appropriate ethical result (H's diminished culpability), without attributing to R any responsibility. But, in a manner parallel to the case of "robot harm" considered in the previous section, "Harming Robots," this resolution does make essential reference to the cognitive states of R: an example of (human-centered) AI ethics in action. The implications of this move need to be explored in more depth, but it is a promising lead.[11]

[10] A more thorough discussion of this scenario would analyse it in terms of the concept of *meaningful human control*; see, e.g., Santoni de Sio, Filippo, and Jeroen van den Hoven. 2018. "Meaningful Human Control over Autonomous Systems: A Philosophical Account". Frontiers in Robotics and AI. 5.

[11] Compare Diamantis, Mihailis. 2019. "The Extended Corporate Mind: When Corporations Use AI to Break the Law". *SSRN Electronic Journal*.

The second human-centered way of dealing with the situation goes beyond a mere interpretive scheme, instead proposing an extension to the designs used in the logical ethical robot methodology. It is proposed that the formalism (such as the one from Scheutz and Malle, displayed earlier) be extended in two ways:

1. Obligations, permissions should be explicitly relativized to the subject to which they apply (i.e., by making O and ◊ a relation between propositions and variables that range over human subjects).
2. Obligations and permissions should be capable of explicitly depending on the cognitive states of subjects (and perhaps the AI technology employed by those subjects, if this second approach is being combined with the technologically extended epistemic states solution, proposed earlier).

On this approach, R would not, *per impossibile*, reason about its own obligations and permissions (it has none), but would instead reason about whether its actions are compatible with H's obligations and permissions. This will not only allow R to derive the genuinely ethically best course of action, but it will also facilitate analysis to correctly allocate responsibility.

Conclusion

In this chapter, I hope to have shown how a human-centered approach can resolve some problems in AI and robot ethics that arise from the fact that (current and foreseeble) AI systems and robots have cognitive states, and yet have no welfare, and are not responsible. In particular, the approach allows that violence toward robots can be wrong even if robots can't be harmed. Also, the approach encourages us to shift away from designing robots as if they were human ethical deliberators. Rather, the slogan goes:

Don't seek to build ethical robots; seek to build robots ethically.

It was found that the cognitive states of AI systems and robots may have a role to play in the proper ethical analysis of situations involving them, even if it is not by virtue of conferring welfare or responsibilities on those systems or robots. Even if robots lack welfare, their cognitive/informational states might make them sufficiently resemble humans to render them unacceptable targets for violence. Even if robots cannot be responsible, their cognitive/informational states may be relevant when assessing the culpability/mens rea of humans interacting with them.

Bibliography

Anderson, Michael, and Susan Leigh Anderson. 2011. *Machine Ethics*. New York: Cambridge University Press.

Arkin, Ronald C. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, Fla: Chapman & Hall/CRC Press.

Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence,* ed. Keith Frankish and William M. Ramsey, 316–34. Cambridge: Cambridge University Press, 2014.

Coeckelbergh, Mark. 2019. *Introduction to Philosophy of Technology*. New York: Oxford University Press.

Lin, Patrick, Keith Abney, and George A. Bekey. 2012. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, Mass: MIT Press.

Müller, Vincent C. "Ethics of artificial intelligence and robotics," in *Stanford Encyclopedia of Philosophy, ed.* Edward N. Zalta. Palo Alto: CSLI, Stanford University, 2020 (forthcoming)

To appear in Markus Dubber, Frank Pasquale, and Sunit Das (Eds.), *Oxford Handbook of Ethics of AI,* Oxford University Press, 2020.

Sparrow, Robert. "Killer robots," *Journal of Applied Philosophy* 24 (1): 62-77, 2007.

Sparrow, Robert. "The march of the robot dogs," *Ethics and Information Technology* 4 (4): 305-318, 2002.

Wallach, Wendell, and Colin Allen. 2010. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Whitby, Blay. 1996. *Reflections on Artificial Intelligence: The Legal, Moral and Ethical Dimensions*. Exeter: Intellect Books.