INVITED PAPER

Motivations and Risks of Machine Ethics

By Stephen Cave, Rune Nyrup, Karina Vold, and Adrian Weller

ABSTRACT | This paper surveys reasons for and against pursuing the field of machine ethics, understood as research aiming to build "ethical machines." We clarify the nature of this goal, why it is worth pursuing, and the risks involved in its pursuit. First, we survey and clarify some of the philosophical issues surrounding the concept of an "ethical machine" and the aims of machine ethics. Second, we argue that while there are good prima facie reasons for pursuing machine ethics, including the potential to improve the ethical alignment of both humans and machines, there are also potential risks that must be considered. Third, we survey these potential risks and point to where research should be devoted to clarifying and managing potential risks. We conclude by making some recommendations about the questions that future work could address.

KEYWORDS | Ethical alignment; ethical reasoning; machine agency; machine ethics

I. INTRODUCTION

Machine ethics is a research field which studies the creation of "ethical machines." This paper aims to clarify what the project of building "ethical machines" amounts to, why it is a goal worth pursuing, and the risks involved. Questions about motivations and risks are important for any field of research. As there are only limited resources for scientific research, utilizing these in an ethically responsible manner requires clarity about what a given field of research aims to achieve, whether this is a desirable and feasible goal, and whether it involves any serious risks, either to researchers, users, or the wider public [1], [2].

Specifically, this paper aims to make three contributions. We start by surveying some of the underlying philosophical complexities involved in the concept of an "ethical machine." We then outline some potential benefits that the creation of such machines may bring and potential risks associated with this research agenda. We conclude that, alongside the positive project of creating ethical machines, more research should be devoted to clarifying and managing potential risks. As we highlight throughout, our aim in surveying these questions is primarily to identify potential problems or complexities that future research might address, rather than to resolve them in this paper.

A few notes on terminology: First, in this paper we use the term "machine" in the broadest sense, to include (among others) both ordinary physical machines, autonomous robots, as well as purely algorithmic systems. So lawn mowers, ATMs, cleaning robots and smartphone apps are all included. However, our main concerns in this paper focus on a specific kind of machines, namely those with a capacity for "ethical reasoning." Thus, many of the above examples, in their present form, are not our primary interest. However, as they will sometimes provide illuminating limiting cases, we still use "machine" in the broad sense and qualify the term appropriately when we have the narrower subset in mind.

Second, we restrict the term "machine ethics" to research which directly contributes to the creation of ethical machines. This includes attempts by engineers and scientists to actually build such machines and theoretical research aiming to facilitate or enable this, but not broader philosophical inquiries into the implications of this technology. The latter field, of which this paper is an example, is sometimes called "machine metaethics" [3].

Manuscript received December 12, 2017; revised February 16, 2018, May 10, 2018, July 16, 2018, and July 26, 2018; accepted August 9, 2018. This work was supported in part by the David MacKay Newton Research Fellowship at Darwin College, Cambridge University, in part by the Leverhulme Centre for the Future of Intelligence, Leverhulme Trust, under Grant RC-2015-067, and in part by The Alan Turing Institute under EPSRC under Grant EP/N510129/1 and Grant TU/B/000074. *(Corresponding author: Karina Vold.)*

S. Cave, **R. Nyrup**, and **A. Weller** are with the Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge CB2 1SB, U.K. (e-mail: sjc53@hermes.cam.ac.uk; rn330@hermes.cam.ac.uk; aw665@cam.ac.uk).

K. Vold is with the Leverhulme Centre for the Future of Intelligence, Faculty of Philosophy, University of Cambridge, Cambridge CB2 1SB, U.K. (e-mail: kvv22@cam.ac.uk).

Digital Object Identifier 10.1109/JPROC.2018.2865996

Table 1 Overview of Paper Contents

Section	*Title*
I.	Introduction
II.	What is an "Ethical Machine"? a. Ethically Aligned Machines b. Ethical Reasoning c. Machine "Agency" d. Summary
III.	Motivations for Machine Ethics a. Individual Machine Decisions b. Fit of Machines with Moral System c. Individual Human Decisions d. Fit of Humans with Moral System e. Summary
IV.	Risks of Creating Ethical Machines a. Failure and Corruptibility b. Value Incommensurability, Pluralism and Imperialism c. Creating Moral Patients d. Undermining Responsibility
V.	Conclusion

Third, note that the terms "ethical" and "moral" are often used interchangeably in ordinary discourse, as well as in much of the literature on machine ethics [4], [5]. Some philosophers draw a sharp distinction between the two [6], [7], but there is no single, noncontroversial way to draw such a distinction [6, esp. fn 1]. For the sake of simplicity, we therefore follow ordinary usage and use the two terms interchangeably. When more fine-grained distinctions are needed, these will be introduced explicitly. We provide an overview of the contents of this paper in Table I.

II. WHAT IS AN ''ETHICAL MACHINE?''

In this section, we aim to clarify some key terms. This will map out four salient issues arising from machine ethics which will structure our discussion in the rest of the paper.

The aims of machine ethics have been expressed in a number of ways, including being able to build machines: which qualify as "artificial moral agents" [8], [9]; which can follow ethical principles [10]; which are capable of ethical decision making [10]; which have "an ethical dimension" [3]; or which are capable of doing things that would require morality in humans analogous to one common definition of artificial intelligence [11]. J. H. Moor has introduced a more fine-grained distinction between different kinds of "ethical machines" which machine ethics might pursue [12]. First, "implicit ethical agents" are machines that are constrained to promote ethical behavior, or at least avoid unethical behavior. Second, "explicit ethical agents" are (in some sense) able to represent or reason about ethical categories or principles. Third, a machine counts as a "full ethical agent" if it is comparable in many or most relevant respects to human moral decision-makers.

While these definitions provide a good starting point, we find that some ambiguities remain with regards to both words in the term "ethical agents."

Start with the term "ethical." In "implicitly ethical," the term "ethical" is used to mean "in accordance with ethics": ethical machines in this sense would be those whose behavior is properly aligned with the relevant ethical principle. For instance, the behavior of an ATM should align with the principle that it is wrong to defraud users of the machine. In many contexts, it will be much more contentious exactly which principles a machine ought to follow—as we explore in the following. Ethical machines in this sense contrast with *unethical* or *immoral* machines, e.g., an ATM which is designed to steal the bank details of users. By contrast, in "explicitly ethical," "ethical" is used synonymously with "involving or relating to ethics": the defining features of ethical machines in this sense is that they are able to reason about ethics or ethical principles. These contrast with *amoral* machines, e.g., a car which has built-in safety features, such as a seatbelt, but does not itself reason about what these should be. To distinguish clearly between these two senses of "ethical," we propose to instead distinguish between *ethically aligned machines*, mirroring the terminology of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design [13], i.e., machines that function in a way which is ethically desirable, or at least ethically acceptable, and machines with a capacity for *ethical reasoning*. We will explain these two senses of "ethical" in further detail in the following sections.

Similarly, the notion of "machine agency" and the term "agent" carry with them certain philosophical connotations that we consider unhelpful to build into the aims of machine ethics. Part of the concern here surrounds whether ascribing machines "agency" will lead to problematic positions on the rights of machines and our responsibilities to them, which we will discuss in greater detail in Section IV-C.

A. Ethically Aligned Machines

Whether something counts as an ethically aligned machine, as defined above, depends on what counts as ethically desirable and acceptable. These are, of course, highly contentious concepts within philosophical ethics. Moreover, it is a basic fact of public ethical and political discourse that people disagree about what is ethically desirable and acceptable behavior. There is thus no reason to assume that there will be a single comprehensive answer to the question of what counts as an ethically aligned machine that would be noncontroversial to all philosophical perspectives, let alone the public at large. We review some of the risks this fact raises for the project of machine ethics in Section IV.

Nonetheless, societies are in practice able to reach consensus or compromise positions which individuals are willing to accept as good enough for collective action, even if they disagree in principle. This is often mediated through social and institutional structures, such as courts, voting, mediation, public consultation processes, etc.) which people respect as legitimate ways of resolving conflicts. Political philosophy contains several theories of how

collective compromises can be legitimately achieved in the face of deep and widespread disagreement [14]–[16].

We will not here give a comprehensive review of existing theories of political legitimacy or opine on which of these best applies to machine ethics (this would be a task for further research). For the purpose of our discussion in what follows, we will adopt the following suggestion as a first approximation and as a general guiding idea: ethically aligned machines are those whose behavior adequately preserves, and ideally furthers, the interests and values of the relevant stakeholders in a given context. We distinguish and discuss more specific ways in which machines might further ethical alignment in this sense in Section III.

There are of course many difficult questions lurking in the interpretation of this formulation: What is a "value" or an "interest"? Are all values and interests equally important? Who are the relevant stakeholders, e.g., does it only include humans? What does it mean to "adequately preserve" the values of different stakeholders, given that these will often conflict? Different ethical theories will give diverging answers and we do not here propose any resolution to these questions. Rather, in adopting the above guiding idea we simply hope to give some indication of what an "ethically aligned machine" might involve and to highlight some of the contentious issues that will be relevant to discussions of ethical alignment.

B. Ethical Reasoning

Reasoning, as we will understand it here, is the processing of information in order to produce a solution to a problem. Different kinds of reasoning can be distinguished in terms of the types of problems they address. This allows us to define ethical reasoning as processes that are concerned with solving ethical problems. Ethical reasoning thus defined can be distinguished, e.g., from mathematical reasoning, which addresses mathematical problems, or factual reasoning about the empirical world. It may be difficult to draw a sharp line between these, but for our discussion it is sufficient that we can distinguish substantially ethical problems, e.g. "should I kill the patient to relieve their pain if they ask me to?," from ones that are mostly factual, e.g. "will this quantity of this drug kill this patient"? We will now discuss some further questions raised by this definition of ethical reasoning.

When talking about (a capacity for) ethical reasoning in machines, the question arises whether machines can have a capacity for reasoning at all. In some contexts, "reasoning" is used in a demanding sense, involving one or more specific capacities such as conscious thought, intentionality, a "faculty of reason," an "openness to the world," or understanding of significance. Whether a machine can have such capacities is of course a well-known objection to the program of strong artificial intelligence [17], [18]. On the other hand, "reasoning" is also commonly used, especially within AI research, in a broader sense to mean simply whatever processing is carried out to reach a conclusion. This is the sense employed when ascribing implicit or unconscious inferences to humans or when talking about automated reasoning by computational systems. Some might argue that ethical reasoning, properly understood, can only be reasoning in the more demanding sense, e.g., because ethical reasoning requires an understanding of the significance of the ethical issues at stake. In our view, this is related to the questions of rights and responsibilities that Moor raised under the label of "full ethical agents." However, we can still ask whether machines are capable of ethical reasoning in the second, weaker sense. Since it is usually this question that machine ethicists are interested in, we use the term "ethical reasoning" in this weaker sense henceforth, unless otherwise noted.

Another objection might be that almost any decision could be construed as solving an ethical problem, which would trivialize the notion of ethical reasoning. To illustrate, suppose a machine is programmed to monitor patients via a camera, use this information to infer their blood sugar levels and offer them insulin if their blood sugar levels reach a predefined threshold. This system is clearly making ethically consequential decisions, but does it exhibit ethical reasoning? We regard this as a limiting case which only involves ethical reasoning to an insignificant degree, where the significance of a reasoning process refers to the difficulty of the problem to be solved by the machine, relative to the resources and inputs it has available. Thus, in this example the machine is solving a significant factual problem by inferring blood sugar levels from video inputs, while the ethical "problem" of applying a single, unambiguous decision rule, e.g. "if blood sugar levels of a patient reaches level T, then offer them insulin," is trivial. By contrast, consider a healthcare robot, as described by Anderson et al. [19], that uses supervised learning to infer a rule for when the duty to protect a patient's health should trump the duty not to violate their autonomy by paternalistically administering medicine. Given the input, inferring a decision rule that adequately balances these two prima facie duties against each other presents a significant problem.

Finally, Anderson et al.'s example [19] involves what is sometimes called a "bottom-up" approach [20] to machine ethics, where the machine infers principles from examples. Should we also classify as ethical reasoning "top-down" (typically symbolic or knowledge-based AI) approaches to machine ethics, where a machine is programmed to infer implications of more general ethical principles for particular contexts or to resolve conflicts between multiple prima facie principles? Again, our framework allows us to distinguish this from simply applying a straightforward decision rule, since the machine needs to solve non-trivial problems in order to derive a fully specified principle for action from high-level or multiple, potentially conflicting principles. Whether any sharp boundary can ultimately be drawn between merely applying a decision rule and inferring the implications of a more general principle is not crucial here. The point is simply that we can distinguish

a spectrum from trivial to more significant (i.e., difficult given the input) ethical problems. Our interest here is in the project of building machines capable of solving ethical problems of the more significant kind.

C. Machine "Agency"

In fields such as robotics, machine learning and artificial intelligence, the term "agent" is sometimes used to refer to anything which has a capacity to act or process information. For example, one influential definition of AI is the study of intelligent agents with capacities for perception and action [21]. This is a wider use than in other disciplines, such as philosophy. But not all in computer science are content with this wide usage, and there is a lively debate about what other conditions machines might need to meet to be ascribed agency (see [4], [8], [12], and [22]–[24], for example).

At least in contemporary western philosophy, a standard account of agency requires the capacity for *intentional* actions. An action is considered intentional when it is caused by the agent's intentional mental states, e.g., her beliefs or desires [25], [26]. Intentional actions are distinguished from mere behaviors, which do no presuppose any intentionality. There are at least two different notions of intentionality: 1) a stronger "realist" sense, which is more difficult to attribute to machines; and 2) a weaker "instrumentalist" sense which allows for more straightforward ascriptions. In the realist sense intentional action requires some of the properties we mentioned in relation to the demanding sense of reasoning, such as capacities for understanding and phenomenal consciousness. It is for example, unlikely that a simple machine, such as a Roomba, has a capacity for intentional action in this strong sense because it lacks genuine conscious beliefs and desires, of the sort that humans have. Roombas operate on syntactic and perceptual feature-based categories, rather than semantic categories, and thus they have no understanding of what they are doing [17]. On the second, instrumentalist sense of intention, whether beliefs and desires can be ascribed to an entity depends entirely on how useful such ascriptions are for explaining its behavior [27]–[29]. On this view, if it is useful to ascribe beliefs and desires to the Roomba in order to explain its behavior, then this usefulness is sufficient for doing so. This view makes it more plausible to ascribe intentional agency to machines [4], [28].

This debate is important to philosophers because of a long-standing tradition that intentionality is a defining mark of the mental [17], [30]–[32], and as such attributions of intentional agency are often connected with questions about mental life, including the reasoning processes, consciousness, and free will, of the agent [12]. In addition to intentionality, it is often thought that ethical agents might require some further condition, e.g., the ability to act in a way that shows understanding of responsibility to other agents [4] or the ability to monitor their behavior in light of their ethical duties and the foreseeable harms that their actions may cause [9].

As this brief summary indicates, whether machines can be called ethical agents in any strong sense is a contentious philosophical issue. In our view it is important to ask whether machines have these capacities because of their links to two distinctly ethical issues, namely: i) whether machines have responsibilities; and ii) whether they have rights. Each of these have links to the notion of agency. With respect to i) if a machine were able to understand its duties or the foreseeable harms of its actions, in the realist sense described above, it would be tempting to regard it as responsible for any harms it causes. With respect to ii) some have held that any being that has goals or desires has an ethical status that should be respected [33]. If a machine can be ascribed intentional states, then, this could entail that we have the responsibility to take into account its rights [34], [35]. But all of this seems to presuppose the stronger sense of intentional agency—which it is more difficult to attribute to machines. Exactly *how* difficult, and indeed whether any given current or future AI system or robot would qualify as having intentional agency in this stronger sense, is controversial. We do not intend to defend any particular account of this here. Many philosophers however agree that merely having intentionality in the instrumentalist sense is not sufficient to ground any important rights or responsibilities. What would be sufficient is an open and controversial question.

In our view, the term "ethical agent" will inevitably carry connotations of these complicated debates. Our primary goal in this section has been to highlight complexities, rather than to resolve them. While the questions of whether machines can have important ethical responsibilities or rights are important, and will be discussed more as follows, we think it is unhelpful to build these connotations into the definition of the aims of machine ethics. Going forward we will therefore talk about machines, rather than agents, and will bring in questions of rights and responsibilities separately.

D. Summary

Based on the preceding discussion, we can restate the main issues highlighted by Moor's framework as follows:

1) building machines that are ethically aligned;
2) building machines that have a capacity for ethical reasoning;
3) building machines that have: i) moral responsibilities or ii) rights.

We believe these capture the main issues at stake in the project of trying to build "ethical machines." To the extent that: 1) is possible, this is presumably something any engineering discipline should aim towards: ATMs and price comparison algorithms should be designed so that they do not defraud users, cars so that they have an ethically acceptable level of road safety, and so on. Sometimes, "machine ethics" is used in a broad sense

to mean the project of constructing ethically aligned autonomous machines. Machine ethics in the narrower sense that we are interested in here is distinctive in that it pursues 2) as a means to achieving 1). Perhaps a health care robot will be better ethically aligned if it is able to infer whether reminding patients to take their medicine would be undue paternalism or simply due care? Pursuing 2) in turn raises questions about whether 3.i) or 3.ii) is necessary for, or could be a side-effect, of 2). Is there a point where a capacity for sophisticated moral reasoning would require us to recognize, say, a chatbot as having rights and responsibilities of its own?

The remainder of this paper will examine these issues. First, in Section III, we consider how 2)—ethical reasoning—might contribute to 1)—ethical alignment. Second, in Section IV, we survey some possible risks arising from this project, including potential side-effects involving 3.i)—responsibilities—and 3.ii)—rights.

III. MOTIVATIONS FOR MACHINE ETHICS

This section discusses the ethical motivations for pursuing machine ethics. These rest on the claim that building machines with a capacity for ethical reasoning will further what we call their ethical alignment, and therefore the interests and values of the relevant stakeholders. As already mentioned, there are unresolved problems in characterizing ethical alignment, stemming from the fact of pervasive disagreement. However, for the purposes of this section, we shall set these aside and focus on presenting the positive arguments for pursuing machine ethics. This is not a purely academic exercise; as we pointed out, adequate ways of overcoming deep ethical disagreements already exist in some domains, and it seems entirely possible this could be developed for machine ethics too.

To organize our discussion, we want to start by introducing a framework to distinguish some ways in which giving machines a capacity for ethical reasoning might enhance their ethical alignment. First, we distinguish two ways of improving the ethical alignment of a machine:

a. improving the behavior of the machine itself;
b. improving the behavior of human decision-makers using the machine.

Second, we distinguish two senses in which we can improve the behavior of machine or human decision-makers:

i) improving individual decisions;
ii) improving how decision-makers fit within morality as a broader social system.

On the one hand, i) we can look at an individual decision of a machine or a human and ask whether it aligns with the standards for morally desirable or acceptable behavior (whatever we take these to be). But, on the other hand, ii) we can also evaluate the ethical alignment of a decision-maker in terms of how it relates to and interacts with other decision-makers. Morality is not just a set of standards for the behavior of individuals; it is also a social system where decision-makers rely on and trust each other, where they can give and ask for explanations of why certain actions were taken, and where apologies or reparations can be offered when mistakes are made. As we explain in more detail in Section IV-D, many of the deepest risks arising from machine ethics will concern the question of how machine decision-makers will fit into or change morality as a social system.

Combining these two distinctions gives us a typology of four ways of enhancing ethical alignment. Within each of these, we may further ask what standard of ethical alignment we are aiming for. For instance, we may merely aim to secure ethical alignment to a human-level acceptable standard, i.e., to the standard of what we would find minimally ethically acceptable from a human. Notice that for humans, even this standard is not trivial; human decision-makers often fall below the standards we expect of them, whether through accident, malice or failures of reasoning. But we can also aim for increasingly higher standards of human-level desirable behavior. Furthermore, it has even been suggested by some that machine ethics could improve the ethical alignment of (machine or human) decision-makers beyond currently existing human standards. We say more about what this might mean in the following.

In the rest of this section, we will survey how proponents of creating ethical machines have argued that this might improve alignment according to each of the above four basic categories. We will also consider how this could be achieved according to different standards.

A. Individual Machine Decisions

Most commonly, machine ethics (in our sense of building machines with a capacity for ethical reasoning) is motivated through examples of autonomous systems currently being developed—e.g. self-driving cars, autonomous weapons or health-care robots—which will be making morally consequential decisions. Giving these systems a capacity for ethical reasoning, machine ethicists argue, will help ensure that their decisions are ethically aligned [36]. If so, and if we will inevitably see more and more autonomous systems deployed, this would provide a moral reason for pursuing machine ethics; it will further the interests and values of the relevant stakeholders.

While this motivation is prima facie plausible, it should be noted that there is no necessary connection between moral reasoning and ethically aligned decisions.

Firstly, a capacity for moral reasoning does not, in itself, guarantee ethically aligned decisions, as humans so often demonstrate. Limited computational power, inaccuracies in the premises or training data supplied to machine reasoners, and limits from the nature of ethics itself may prevent a machine from making ethically aligned decisions, even if it has a capacity for moral reasoning [37]. Most machine ethicists would presumably agree that ethical reasoning does not guarantee full ethical alignment. Instead,

the motivation for pursuing machine ethics rests on the more modest claim that enough incremental progress can be made for machine ethics to make a positive contribution to the ethical alignment of machine decision making [38].

Assuming, then, that machine ethics could make a positive contribution to the ethical alignment of machine decisions, we should still ask whether it is necessary, or more precisely, a cost-effective means, compared to other options. There are many contexts where machines without a capacity for ethical reasoning, or machines which only solve trivial ethical problems, already function in ethically unproblematic ways. Examples include factory robots, automated cash machines or automated metro trains [9], [10], [12]. Ethical alignment in these cases is achieved through appropriate safety measures and external constraints on the machines' functioning. For example, to deter fraud cash machines will only dispense cash under specific circumstances, e.g., with a valid card and PIN code, and are furthermore constrained in the amount of cash they can dispense. The decision that these constraints are appropriate for cash machines is of course informed by human ethical reasoning, but it does not require the machine to solve any significant ethical problems; it merely follows these predefined rules.

However, machine ethicists argue that when machines are required to operate with flexibility in a causally complex environment, endowing them with a capacity for moral reasoning becomes important, e.g. [12] and [36]. To evaluate this argument, notice that mere causal complexity—i.e., environments where a wide range of relevant causal factors can combine in an open-ended number of ways—does not always necessitate a capacity for moral reasoning. For example, a sophisticated autonomous system might have to carefully manage highly complex processes within an automated factory. However, if the only morally relevant concern is to ensure that all machines shut down when humans enter the production floor, the system does not need to engage in any significant ethical reasoning. While it might be a complex factual problem to determine whether a human is present, no additional ethical reasoning is required to determine whether to shut down once this is determined. As long as it can reliably recognize humans, designers can implement a fully specific, preprogrammed principle for action.

Rather, it is when causally complex environments produce what we can call moral complexity that a capacity for ethical reasoning becomes important. By "moral complexity" we mean cases where: a) human ethicists cannot formulate a general, fully specific principle for action and b) where, due to causal complexity, decision makers can face an open-ended range of morally distinct situations so that these cannot simply be enumerated in advance. Moral complexity can arise for a number of reasons, including the following.

The most straightforward case is when several prima facie duties compete, such as when deciding whether to prioritize the duty to protect the health of a patient over the duty to respect their autonomy. Since human ethicists require contextual judgments to resolve the dilemmas arising from such situations, this creates moral complexity when the designers of a machine cannot predict in advance the combinations of factors in all such potential situations.

Another form of moral complexity can arise when causal complexity gives rise to uncertainty. In such cases, decision makers might have to balance the risks of false negatives against risks of false positives. Unless all possible situations that a machine may encounter can be specified in advance, this in turn requires comparing the moral "weight" of these risks, whether by assigning numerical utilities or some other means, which plausibly requires some capacity for significant ethical reasoning.

Not all morally complex situations involve conflicting principles per se. For example, determining which principles or procedure should be applied in a given situation requires the ability to determine the ethically relevant aspects of the situation, e.g., whether it involves dilemmas or tradeoffs. Sometimes, a machine may face an environment where causal complexity makes it nontrivial to isolate the morally relevant aspects of the situation. Again, as there is usually no general, fully specific principle for identifying the relevant aspects of a situation, a capacity for this type of ethical reasoning may be conducive to ethical alignment in these cases.

B. Fit of Machines With Moral System

Even if we can secure that the individual decisions of a machine are ethically aligned to a human standard, many would still be hesitant to let it replace human decision-making if it is unable to explain and justify its decisions. Several scholars have highlighted the notion of "explainable AI" as important for creating trustworthy and accountable machines, e.g., [39]–[42]. Exactly what kinds of explanations are required in which contexts is still a matter of debate. However, one important candidate, which machine ethics could plausibly help address, is the ability to explain the ethical reasons behind a decision.

In our terminology, this is an example of how building machines with the ability to explain their reasons for acting could improve their fit with morality as a social system. It is an important part of human moral social systems that we are able to give and ask for the reasons behind our decisions, including in particular our ethical reasons such as which values we prioritized in a given situation. For example, we might prioritize preventing harm in certain scenarios, or respecting autonomy in others. Explaining why we thought a given decision was morally justified, even if only retrospectively, allows us to challenge each other, become convinced that the other person was right after all, or to demand apologies or reparations when the reasons given are unsatisfactory. To participate in these aspects of our moral systems, machines would need a capacity to represent and communicate moral reasoning in a format understandable to humans. Furthermore, if these explanations are to be anything more than "just-so"

stories, they should at least to some degree reflect the actual reasoning processes behind the machine's decisions.

Having the capacity to explain its reasons is arguably not sufficient for a machine to participate adequately in these social aspects of morality. For example, since machines lack a capacity to feel remorse, some might question whether they can offer genuine apologies. Notice, however, that group agents, such as states or corporations, do sometimes seem to apologize for their actions, although they presumably do not have feelings of remorse either. This may be explained by the fact that group agents have other relevant attributes which machines lack, e.g., they own property that can be offered as reparations, and they are constituted by human agents, who may feel remorse on behalf of the group agent.

We will not pursue the analogy between group agents and machines further here (on group agency in general, see [43]–[46]). Our more general point is this: while not sufficient in itself, it may be possible to situate machines with a capacity for ethical reasoning within a broader social or legal framework which makes it possible for apologies to be offered (whether by the machine, its owners, or its designers). In this case, a capacity for ethical reasoning would be conducive to enhancing the ethical alignment of machines by virtue of improving its fit with these system-level aspects of human morality.

The ability to give reason-based explanations becomes especially important if we require autonomous systems to operate in morally complex domains where human ethicists cannot formulate clear outcomes-based means of monitoring their performance. For example, a medical AI-system that is only tasked with recommending patients for a specific cancer treatment is relatively straightforward to evaluate: if the system leads to a decrease in mortality and morbidity both from untreated tumors and from unnecessary treatments, we would arguably have good, outcome-based reasons to trust it, even if the system cannot explain its reasoning. By contrast, a system in charge of managing all treatment decisions within a hospital would have to make many ethically contentious decisions about, e.g., which patient groups should be prioritized. Since the trustworthiness of a machine in such cases will to some extent rely on the reasons they can give for its actions, we may very well require machines to satisfy above-human level standards of explainability. To the extent that we want autonomous systems to operate in such contexts, it becomes all the more important that they can accurately represent and communicate the moral reasoning behind their actions.

C. Individual Human Decisions

Some proponents of building machines with a capacity for ethical reasoning argue that doing so might also improve the ethical alignment of humans. The first way it might do this is by improving individual human decisions.

Human reasoning is prone to a number of imperfections that we rightly regard as failings: our decisions are influenced by biases, self-interest and sloppy reasoning. We, as humans, are worryingly adept at fooling ourselves and others into thinking that our actions are morally sound. Machine ethicists have argued that an automated moral reasoner will be free from many of these human limitations [8], [47]–[49]. Suppose that we could build a moral reasoning system able to compute and highlight to us the implications of our moral commitments, say, that my commitment to mitigating climate change is inconsistent with ordering beef steak at the restaurant. Even if such a system were not implemented into any autonomous agent, it might still be able to improve and extend human moral reasoning, analogous to the way pocket calculators improve and extend human numerical reasoning.

D. Fit of Humans With Moral System

In addition to improving our individual decisions, S. L. Anderson [50] furthermore argues that machine ethics might help improve human morality as a whole, by helping us to formulate clearer, more consistent ethical theories and to achieve increased consensus on moral dilemmas. For instance, Anderson argues that if philosophical ethicists try to formulate their theories in a format that can be computed by a machine, this would force them to face the implications of their theories squarely. To improve ethical theorizing in this way would arguably require machines capable of representing moral reasoning explicitly. It should be able to reveal to philosophical ethicists not just what the implications of a given ethical theory are, but how conclusions are reached.

Some ways of improving human morality might aim to ensure our actions consistently meet our current standards. For instance, Anderson and Anderson's ethical guidance systems learn to resolve ethical dilemmas based on training examples where human ethicists agree on the right resolution [19]. Such a system might be able to raise human morality to the level of the existing consensus of expert human ethicists. Other implementations of machine ethics may promise to go beyond the current consensus and, e.g., resolve outstanding moral disagreements or uncover where the current consensus could be improved. Some proponents of machine ethics suggest that it might thereby be able to actively promote human moral progress [49].

Even if machine ethical reasoning does not allow humans to reach increased consensus, e.g., if some disagreements are fundamentally unresolvable (cf. Section IV-B), they may still improve the fit of humans within moral systems by helping to explain and make comprehensible those disagreements. An improved ability to understand and explain the nature of our disagreements to each other could conceivably improve our ability to negotiate or otherwise manage such conflicts.

E. Summary

While some of the potential benefits outlined above are mostly speculative promises at this stage, the potential

benefits are large enough to provide a prima facie motivation for trying to build machines with a capacity for ethical reasoning. Crucially, however, these benefits need to be weighed against any potential risks that might either be inherent in achieving ethical alignment through machine ethics, or that might arise as by-products. We survey some of the most salient such risks in the next section.

IV. RISKS OF CREATING ETHICAL MACHINES

In this section, we survey four broad categories of risks: A) the risk that ethically aligned machines could fail, or be turned into unethical ones; B) the risk that ethically aligned machines might marginalize alternative value systems; C) the risk of creating artificial moral patients; and D) the risk that our use of moral machines will diminish our own human moral agency.

A. Failure and Corruptibility

As we mentioned before, having the capacity for moral reasoning does not guarantee ethically aligned decision making. Charging machines with ethically important decisions thus carries the risk of reaching morally unacceptable conclusions that would have been recognized as such by humans.

First, even the best reasoner can reach false conclusions if they rely on false premises. The simplest case of this is if a machine relies on misleading information about the situations it acts in, say, if it fails to detect that there are humans present which it ought to protect. Relatedly, some have highlighted that the computational intractability of predicting the effects of acting in complex social situations might lead even an infallible moral reasoner with perfect information to ethically unacceptable conclusions [8], [37]. Furthermore, if the moral principles or training examples that human developers supply to a system contain imperfections, this may lead to the robot inferring morally unacceptable principles [37].

While it would be an important milestone for machine ethics to be able to ensure that a machine makes moral decisions at minimally acceptable human standards, this may not be good enough for machines. First, while we might accept some mistakes from individual humans, if an autonomous system is applied on a global scale, such as an autonomous vehicle from a large manufacturer, individually minor but systematic mistakes may amount to very serious problems in the aggregate. Second, we may accept certain levels of average reliability from humans because we have developed ways to predict and manage those mistakes. However, as illustrated by examples of adversarial techniques in machine learning [51], machines can often fail in ways different to humans—i.e., they may be liable to fail under circumstances where humans would usually not fail, or they may produce different kinds of errors than humans when they fail. Thus, the risk is that when machines fail, they do so in ways that are difficult to predict or manage. Exactly what levels of performance would be acceptable for a machine is currently not clear and likely to be context specific.

A further risk arises from the possibility of moral reasoning systems being easily corruptible [52], [53], whether by malicious designers, hackers or coding errors. This risk would be further compounded if malicious machines at the same time had a powerful capacity for producing deceptive or manipulative explanations for their actions. Machines with a capacity to produce convincing ethical explanations might be exploited to convince humans to accept serious divergences from ethical alignment.

To be sure, many currently existing machines without the capacity for ethical reasoning are also vulnerable to error and corruptibility. Shying away from building machines with ethical reasoning will not solve this problem. It is possible that incorporating ethical reasoning into existing systems can make them more resilient to these problems. However, our concern here is that ethical reasoning capacities may themselves be vulnerable to error and corruptibility—perhaps even especially vulnerable, as Vanderelst and Winfield argue [52]. If the very same technique that would give machines the capacity for moral reasoning can easily fail or be corrupted to produce unethical behavior, this would provide a severe counter-weight to any positive reasons for pursuing machine ethics. At the very least, care should be taken not to reintroduce the problems that machine ethics was supposed to solve.

B. Value Incommensurability, Pluralism, and Imperialism

The circumstances discussed in the previous section were ones in which there were: a) definite facts as to what a morally correct outcome or action would be but b) risks that the morally correct outcome or action might not be pursued by the automated system for one reason or another. There may, however, be circumstances in which there is no definite fact about what is morally correct. Here we discuss the risks associated with automated moral decision making in these contexts.

Value pluralism maintains that there are many different moral values, where "value" is understood broadly to include duties, goods, virtues, or so on [54]. If value pluralism is true, then we cannot reduce all values to one single value, such as happiness or pleasure. Value monists deny that there could ever be such circumstances, instead maintaining that there is always a definite fact about how to act morally, or what the best outcome is. I. Kant, for example, defended the view that there is one moral principle that moral agents should abide by, and that any other moral principles could be reduced to that one [55]. Not all deontologists agree. As a value pluralist, W. D. Ross [56] thinks that there is a multitude of moral duties that may sometimes conflict. Furthermore when duties conflict, the dilemma may be genuinely irresolvable. Unlike monists, value pluralists tend to believe that there are at least some,

perhaps many, complex moral dilemmas that result from a conflict between competing and incommensurable values and which cannot be resolved, e.g. [57], [63].

To put the distinction in mathematical terms, monists claim that there is a total ordering on the set of all possible actions, while value pluralists claim that there is only a partial ordering. If the latter is the case, we cannot expect a machine to be able to resolve such dilemmas as no such solution would exist [3], [24], [37].

Furthermore, some argue that there are reasons to preserve this plurality or diversity. When confronted with moral dilemmas that have no resolution, humans must sometimes act. In these cases the action or outcome may be unsatisfactory. However, as we argue above, most humans have a limited sphere of influence, but the same may not be true for machines that could be deployed en masse, while governed by a single algorithm. Thus whatever heuristic is employed to overcome any genuinely irresolvable dilemmas could be highly influential. This could result in something akin to value imperialism, i.e., the universalization of a set of values in a way that reflects the value system of one group (such as the programmers). This could be pursued intentionally or, perhaps more alarmingly, could also be perpetrated inadvertently if programmers unintentionally embed their values in an algorithm that comes to have widespread influence. Such value imperialism might affect (or disrupt) cultures differently, or degrade cultural autonomy.

C. Creating Moral Patients

Earlier we explained that machines created to have their own ethical reasoning capacities could also ipso facto have attributes we associate with genuine agency. We also flagged some of the philosophical issues that arise from attributing genuine agency to machines, and pointed out how talk of machines as agents is becoming commonplace (again, see [4], [5], [8], [22]–[24], and [64]). Somewhat paradoxically, while machine ethicists may be pursuing the moral imperative of building machines that promote ethically aligned decisions and improve human morality, doing so may result in us treating these machines as intentional agents, which in turn may lead to our granting them status as moral patients. We argue that this runs the risk of creating new moral duties for humans, duties that may constrain us in important ways and expand our own moral responsibilities.

We noted before that humans are both moral agents and moral patients. Our moral responsibilities stem from our agency: because of our ability knowingly to act in compliance with, or in violation of, moral norms we are held responsible for our actions (or failures to act). At the same time, we are also moral patients: we have rights, our interests are usually thought to matter, and ethicists agree we should not be wronged or harmed without reasonable justifications.

These two concepts—moral agency and moral patiency—can be clearly separated, but they might nonetheless be interrelated in practice. So whereas moral agency is not necessary for status as a moral patient (for example, we might consider babies or some animals to be moral patients, but not moral agents), it might be sufficient. That is, the very capacities that underpin moral agency might also justify a claim to moral patiency. If that were the case, then by creating artificial moral agents, we may (unintentionally) create moral patients.

What grounds moral patiency is much debated. But the modern view, defended by many philosophers today, points to sophisticated cognitive capacities [65]. Different candidate capacities have been defended, for example, the capacity to will [66] or the capacity for some kind of self-awareness [67], [68]. This tradition of pointing to different intellectual capacities goes back at least as far as Kant [55]. Another kind of cognitive capacity that is often posited as sufficient for moral status (or at least some degree of moral status, for those who allow that moral status admits of degrees) is the ability to feel pain or to suffer.

While they might one day, it seems unlikely that current machines have developed phenomenal consciousness, and thus, unlikely that they feel pleasure or pain, or have the capacity to suffer [22], [69]. More likely, however, is that machines will possess other sophisticated reasoning capacities that might lead us to treat them as moral patients. As mentioned, for some, self-awareness (or self-monitoring), the ability to reflexively represent oneself, grounds moral status. Although currently most or all machines also lack this capacity, there already exist some reasonable exceptions. For example, some algorithms operate with hierarchical layers of neural networks, where higher levels predict the probability of success for lower layers, thereby engaging in a kind of self-monitoring and self-representation [69], [70]. Kant would have us ask whether or not machines are capable of their own, autonomous, practical reasoning, in which case this could ground their dignity and require that we not treat them as mere means to our ends. We have already seen that building machines with autonomous moral reasoning capacities is the explicit aim of, and grounds the moral motivation for, machines ethics.

There is significant risk in building machines that would qualify as moral patients. As responsible moral agents, we would be obliged to take their interests seriously. This could potentially have huge costs: we might not be able to use such machines as mere tools or slaves, but might have to respect their autonomy, for example, or their right to exist and not be switched off. If we had reached a point where our economy, and systems from healthcare to education, depended significantly on AI, this could be hugely disruptive. We might also have to share our privileges: for example, by giving suitably advanced AI's a right to vote, or even a homeland of their own. Consequently, Bryson [71] has argued that engineers have a responsibility not to build sentient robots, so we do not have any special obligations to them. She argues that robots should be our "slaves" and should serve us without our owing them anything

(though tools might be a better analogy, as most would now recognize that slaves were unjustly denied the status of moral patients).

D. Undermining Responsibility

A fourth potential risk of machine ethics is that it will undermine human moral agency—that is, it will undermine our own capacity to make moral judgements, or our willingness and ability to use that capacity, or our willingness and ability to take responsibility for moral decisions and outcomes.

Such cases could arise as a result of what is known as the "automation paradox," a general problem which arises for most labor-saving machines. In this section we show how this problem applies to machines capable of ethical reasoning and highlight the ethical challenges this raises.

Harford [72], [73] identifies three strands to this problem: 1) automated systems "accommodate incompetence" by automatically correcting mistakes. 2) Even when the relevant humans are sufficiently skilled, their skills will be eroded as they are not exercised. And 3) automated systems tend to fail in particularly unusual, difficult or complex situations, with the result that the need for a human to intervene is likely to arise in the most testing situations, for which the human might be ill-prepared.

All three of these strands have direct bearing on moral decision making. The first strand could be relevant to circumstances in which either the goal of the automated systems was for a machine to make ethical decisions alone, or if its goal was to assist a human in making such decisions. In the first case, where the machines are making the decisions, it is possible that humans in the environment would consequently not develop the relevant skills themselves. For example, in the case of the healthcare robot, human staff might not develop the requisite judgment and sensitivity to decide when to intervene paternalistically to ensure a patient took their medicine. In cases where the human is making the decision, it is possible that machine-assistance would ensure that deficiencies in the human's own moral reasoning capacities did not (in standard cases) come to light, in the way that GPS-navigational assistants ensure that deficiencies in a human's own navigational skills do not come to light—except when the system fails.

The second strand of the automation paradox, the risk of skill erosion is also relevant, particularly in cases where the decision-making process is entirely automated (including cases where the system is intended to function at better-than-human level). That moral reasoning is indeed a skill is evidenced by the extent to which it features in the socialization and education of children, and by the fact that it is part of professional education, e.g., in medicine. If we think a lack of practice due to automation can lead to skill-erosion in some settings—Harford cites the case of Air France Flight 447, which crashed after the pilot and co-pilots responded poorly to the plane stalling—there is prima facie reason to think it might do so with regard to moral decision-making.

The third strand compounds the first two. We can imagine machines that successfully navigate the everyday ethical questions and tradeoffs of their environment, such as in a hospital or on the road. We should hope that these machines would also be able to recognize their own limitations, and would alert a human when they encounter a situation that exceeds their training or programming. But there is a good chance that these situations (or some of them) will be more novel or complex than the average, and might therefore be just those that would be more challenging for a human. It is also possible that these decisions will need to be made quickly, for example, in the case of trolley-problem type decisions for autonomous cars, in which either of two possible options will cause significant harm, this would be fractions of a second. This raises the possibility that ill-prepared humans, whose skills have either (strand one) not been developed or (strand two) have been eroded, will have thrust upon them, potentially at very short notice, exactly those moral decisions that are most difficult. It is easy to see how this could go badly.

These problems could be exacerbated if the moral agency of machines increases. As noted above, agency is closely related to responsibility: those entities we tend to regard as full moral agents (healthy adult humans) are those entities we hold responsible for their actions. If machines increase in agency, we will therefore be increasingly tempted to hold them responsible for their decisions and actions, whereas until that point we might have assigned responsibility to human developers, owners or users. There may well be frameworks within which this could go well. But we can also imagine that formal assignation of moral responsibility to machines would exacerbate the automation paradox risks noted above, as humans effectively feel "off the hook" [74].

As the machines and the ethical situations they confront become more sophisticated and complex, these challenges could be exacerbated still further. We noted above that some decisions can range over a wide range of values and other variables (for example, prioritizing resources in a hospital), such that we find it difficult to rely only on outcomes-based means of monitoring, and instead rely also on the broader moral system of explanation and reason-giving. For some classes of algorithm, interpretability/explicability of this kind already poses significant technical challenges [40], [75]. But it is possible that this could come to pose a challenge for any system.

Our human system of reason-giving is of course based on what we humans can understand. For machine decisions to be understandable similarly means understandable to us humans, with our particular cognitive capacities and limitations. It is conceivable that ethical machines have the potential to make decisions in domains whose complexity exceeds our human capacities to understand, for example, where very many lives are affected in different ways over long timescales, requiring large numbers of tradeoffs. In such cases, the notions of reason-giving, transparency and interpretability are severely challenged. Perhaps a suitably

sophisticated machine could attempt to communicate its reasoning to us, but only by grossly simplifying, in the way that an adult human might simplify a moral argument for a small child. But it is difficult to see how humans could meaningfully hold a machine to account through the system of reason-giving in such circumstances.

We can imagine the extreme case: ethical machines are deployed increasingly in everyday settings, from care homes to schools, and perform well at easily understandable goals. More sophisticated systems are then used to advise on more complex matters. After establishing a track record, decision-makers, from police officers to town-planners, come to rely on these systems, increasingly delegating decisions to them. Improvements in many aspects of private and public life are widely enjoyed and credited to the machines. They are therefore tasked with intervening in domains at levels of sophistication that exceed human capabilities, whether it be improving traffic flow or improving the human genome. Again, benefits are enjoyed, but no human is able any longer to understand what the machines are doing.

In such a case, the humans would be well on the way to abdicating moral responsibility for decisions made on their behalf. Some might consider this worth whatever good outcomes might be enjoyed as a result of the machines' actions. Of course, such a scenario would bring with it all the risks of the automation paradox, creating considerable hazard should the machines fail. But it also brings an additional, more disturbing worry: as the humans in this scenario cease to use their moral faculties, the risk increases that they would not even know what it meant for the machines to fail. Passing enough consequential ethical decisions over to machines too complex for us to understand could therefore pose a risk to the entire system of moral reasoning, reason-giving and responsibility.

V. CONCLUSION

In this paper, we have tried to clarify the aims and risks of creating ethical machines. We have argued that there are good prima facie reasons for pursuing this research agenda. As argued in Section III, designing machines with a capacity for moral reasoning could potentially improve the ethical alignment of both humans and machines. However, these prima facie reasons do not in themselves give sufficient reason to pursue machine ethics unless the risks highlighted in Section IV can be properly managed, either by developing solutions that can mitigate the risks when they arise or by formulating regulations restricting the use of automated moral reasoning to low-risk contexts.

A crucial first step is therefore to obtain additional clarity on whether and when these risks are likely to arise. In this paper we have identified the following four themes that future research would need to address.

1) Under what conditions is a given moral reasoning system likely to enhance the ethical alignment of machines and, more importantly, under what conditions are such systems likely to fail? Even if a capacity for ethical reasoning could be shown to have the potential for significantly enhancing the ethical alignment of a machine, this would have to be weighed against the risks of systematic or unpredictable failures. In addition, how can we prevent machine ethical reasoning from being corruptible or employed for malicious, deceptive or manipulative ends?

(2) How do we ensure that such machines are able to adequately deal with value pluralism and deep disagreements? On the one hand, being able to reconcile such disagreements is one of the potential benefits of a machine ethical reasoning system. However, as we have argued, we would not want the machine to rule out benign ethical pluralism by default, e.g., by assuming that there is single, definite answer to all ethical problems.

(3) Under what conditions would we believe we ought to grant moral rights to machines? Would the prerequisites of moral agency fulfill also the conditions of moral patiency? What consequences would it have for us to acknowledge the moral patiency of (suitably advanced) machines?

(4) How can we avoid automated ethical reasoning undermining our moral responsibility? Specifically, what impact might reliance on ethical machines have on our own moral judgement in different sectors and settings? How can we preserve moral autonomy and oversight where machines are making moral judgements in scenarios of increasing complexity?

Some recent work has started to address these issues to some degree, especially the third theme, regarding whether machines should have moral right (e.g. [9], [20], [22], [23], [71], [72]), but still more work needs to be done. □

Acknowledgments

The authors would like to thank H. Price, A. Alexandrova, S. John, J. Hernández-Orallo, K. Dihal, and H. Shevlin for valuable input and discussion on this paper.

REFERENCES

[1] H. Lacey, *Is Science Value-Free?* London, U.K.: Routledge, 1999.
[2] P. Kitcher, *Science in a Democratic Society*. New York, NY, USA: Prometheus Books, 2011.
[3] S. L. Anderson, "Machine metaethics," in *Machine Ethics*, M. Anderson and S. Anderson, Eds. New York, NY, USA: Cambridge Univ. Press, 2011, pp. 21–27.
[4] J. Sullins, "When is a robot a moral agent?" *Int. Rev. Inf. Ethics*, vol. 6, pp. 23–30, Dec. 2006.
[5] R. Tonkens, "A challenge for machine ethics," *Minds Mach.*, vol. 19, no. 3, pp. 421–438, 2009.
[6] G. W. F. Hegel, *Elements of the Philosophy of Right*, A. W. Wood, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1991.
[7] J. Annas, "Ancient ethics and modern morality," *Philos. Perspectives*, vol. 6, pp. 119–136, Jan. 1992.
[8] C. Allen, G. Varner, and J. Zinser, "Prolegomena to any future artificial moral agent," *J. Exp. Theor. Artif. Intell.*, vol. 12, no. 3, pp. 251–261, 2000.
[9] C. Allen and W. Wallach, *Moral Machines: Teaching Robots Right from Wrong*. London, U.K.: Oxford Univ. Press, 2009.
[10] M. Anderson and S. L. Anderson, "Guest editors' introduction: Machine ethics," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 10–11, Jul. 2006.

[11] G. D. Crnkovic and B. Çürüklü, "Robots: Ethical by design," *Ethics Inf. Technol.*, vol. 14, no. 1, pp. 61–71, 2012.
[12] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 18–21, Jul. 2006.
[13] *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being With Autonomous and Intelligent Systems, Version 2*. 2017. [Online]. Available: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
[14] J. Rawls, *Political Liberalism*. New York, NY, USA: Columbia Univ. Press, 1993.
[15] F. Peter, "Political legitimacy," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. 2017. [Online]. Available: https://plato.stanford.edu/archives/sum2017/entries/legitimacy/
[16] R. Binns, "Algorithmic accountability and public reason," *Philos. Technol.*, 2017. [Online]. Available: https://doi.org/10.1007/s13347-017-0263-5
[17] J. Searle, "Minds, brains, and programs," *Behav. Brain Sci.*, vol. 3, no. 3, pp. 417–424, 1980.
[18] H. Dreyfuss, "Why heideggerian AI failed and how fixing it would require making it more heideggerian," *Philos. Psychol.*, vol. 20, no. 2, pp. 247–268, 2007.
[19] M. Anderson, S. L. Anderson, and C. Armen, "An approach to computing ethics," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 56–63, Jul. 2006.
[20] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics Inf. Technol.*, vol. 7, no. 3, pp. 149–155, 2005.
[21] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. London, U.K.: Pearson, 2010.
[22] K. E. Himma, "Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?" *Ethics Inf. Technol.*, vol. 11, no. 1, pp. 19–29, 2009.
[23] L. Floridi and J. W. Sanders, "On the morality of artificial agents," *Minds Mach.*, vol. 14, no. 3, pp. 349–379, 2004.
[24] M. Anderson, S. L. Anderson, and C. Armen, "Towards machine ethics," in *Proc. IAAA Workshop Agent Org. Theory Pract.*, San Jose, CA, USA, Jul. 2004, pp. 1–7.
[25] G. E. M. Anscombe, *Intention*. Oxford, U.K.: Basil Blackwell, 1957.
[26] D. Davidson, *Essays on Actions and Events*. Oxford, U.K.: Clarendon Press, 1980.
[27] M. Schlosser, "Agency," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., 2015. [Online]. Available: https://plato.stanford.edu/entries/agency/
[28] D. C. Dennett, *The Intentional Stance*. Cambridge, MA, USA: MIT Press, 1987.
[29] D. C. Dennett, *Brainchildren: Essays on Designing Minds*. Cambridge, MA, USA: MIT Press, 1998.
[30] J. Searle, *sIntentionality: An Essay in the Philosophy of Mind*. Cambridge, U.K.: Cambridge Univ. Press, 1983.
[31] F. Brentano, *Psychology From an Empirical Standpoint*. London, U.K.: Routledge, 1995.
[32] T. Crane, *Intentionality as the Mark of the Mental, Contemporary Issues in the Philosophy of Mind*, A. O'Hear, Ed., 1998.
[33] M. Dawkins, *Why Animals Matter*. London, U.K.: Oxford Univ. Press, 2012.
[34] C. B. Jaeger and D. T. Levin, "If Asimo thinks, does Roomba feel the legal implications of attributing agency to technology," *J. Hum.-Robot Interact.*, vol. 5, no. 3, pp. 3–25, 2016.
[35] T. Wu, "Machine speech," *Univ. Pennsylvania Law Rev.*, vol. 161, pp. 1495–1533, 2013.
[36] C. Allen, W. Wallach, and I. Smit, "Why machine ethics?" *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 12–17, Jul. 2006.
[37] M. Brundage, "Limitations and risks of machine ethics," *J. Exp. Theor. Artif. Intell.*, vol. 26, no. 3, pp. 355–372, 2014.
[38] T. M. Powers, "Incremental machine ethics," *IEEE Robot. Automat. Mag.*, vol. 18, no. 1, pp. 51–58, Mar. 2011.
[39] J. Bryson and A. Winfield, "Standardizing ethical design for artificial intelligence and autonomous systems," *Computer*, vol. 50, no. 5, pp. 116–119, May 2017.
[40] A. Weller, "Challenges for transparency," in *Proc. ICML Workshop Hum. Interpretability Mach. Learn. (WHI)*, Sydney, NSW, Australia, 2017. [Online]. Available: https://arxiv.org/abs/1708.01870v1
[41] K. Baum, M. E. Köhl, and Schmidt, "Two challenges for CI trustworthiness and how to address them," in *Proc. 1st Workshop Explainable Comput. Intell.*, Santiago de Compostela, Spain, Sep. 2017, pp. 1–5.
[42] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Havard J. Law Technol.*, 2017. [Online]. Available: https://dx.doi.org/10.2139/ssrn.3063289
[43] P. Pettit, "Groups with minds of their own," in *Socializing Metaphysics: Nature Social Reality*, F. F. Schmitt, Ed. Lanham, MD, USA: Rowman & Littlefield, 2003, pp. 167–193.
[44] P. Pettit, "Akrasia, Collective and individual," in *Weakness of the Will and Practical Irrationality*, S. Stroud and C. Tappolet, Eds. London, U.K.: Oxford Univ. Press, 2003, pp. 68–96.
[45] P. Pettit, "Rationality, reasoning and group agency," *Dialectica*, vol. 61, no. 4, pp. 495–519, 2007.
[46] C. List and P. Pettit, *Group Agency: The Possibility, Design, and Status of Corporate Persons*. London, U.K.: Oxford Univ. Press, 2011.
[47] H. Seville and D. G. Field, "What can AI do for ethics" *AISB Quart.*, vol. 104, M. Anderson and S. Anderson, Eds. New York, NY, USA: Cambridge Univ. Press, 2000, pp. 499–510.
[48] M. Anderson and S. L. Anderson, "Robot be good," *Sci. Amer.*, vol. 303, no. 4, pp. 72–77, 2010.
[49] G. Marcus. (Nov. 24, 2012). *Moral Machines*. The New Yorker. [Online]. Available: https://www.newyorker.com/news/news-desk/moral-machines
[50] S. L. Anderson, "How machines might help us achieve breakthroughs in ethical theory and inspire us to behave better," in *Machine Ethics*, M. Anderson and S. Anderson, Eds. New York, NY, USA: Cambridge Univ. Press, 2011, pp. 524–530.
[51] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok (2017). "Synthesizing robust adversarial examples." [Online]. Available: https://arxiv.org/abs/1707.07397
[52] D. Vanderelst and A. Winfield (2016). "The dark side of ethical robots." [Online]. Available: https://arxiv.org/abs/1606.02583
[53] V. Charsi (2017). "Towards moral autonomous systems." [Online]. Available: https://arxiv.org/abs/1703.04741
[54] E. Mason, "Value pluralism," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. 2015. [Online]. Available: https://plato.stanford.edu/archives/sum2015/entries/value-pluralism/
[55] I. Kant, *Groundwork of the Metaphysic of Morals*. 1785.
[56] W. D. Ross, *The Right and the Good*. London, U.K.: Oxford Univ. Press, 1930.
[57] D. Wiggins, "Weakness of will, commensurability, and the objects of deliberation and desire," in *Essays on Aristotle's Ethics*, A. O. Rorty, Ed. Berkeley, CA, USA: Univ. California Press, 1980.
[58] D. Wiggins, "Incommensurability: Four proposals," in *Incommensurability, Incomparability and Practical Reason*, R. Chang, Ed. Cambridge, MA, USA: Harvard Univ. Press, 1997.
[59] B. Williams, "Ethical consistency," in *Problems of the Self*. Cambridge, U.K.: Cambridge Univ. Press, 1973.
[60] B. Williams, *Ethics and the Limits of Philosophy*. Cambridge, MA, USA: Harvard Univ. Press, 1985.
[61] I. Berlin, *The Crooked Timber of Humanity*. New York, NY, USA: Random House, 1991.
[62] M. Stocker, *Plural and Conflicting Values*. Oxford, U.K: Clarendon, 1990.
[63] M. Stocker, "Abstract and concrete value: Plurality, conflict and maximization," in *Incommensurability, Incomparability and Practical Reason*, R. Chang, Ed. Cambridge, MA, USA: Harvard Univ. Press, 1997.
[64] in *Proc. Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*. [Online]. Available: http://celweb.vuse.vanderbilt.edu/aamas18/
[65] A. Jaworska and J. Tannenbaum, "The grounds of moral status," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. 2017. [Online]. Available: https://plato.stanford.edu/archives/fall2017/entries/grounds-moral-status/
[66] W. Quinn, "Abortion: Identity and loss," *Philos. Public Affairs*, vol. 13, no. 1, pp. 24–54, 1984.
[67] J. McMahan, *The Ethics of Killing: Problems at the Margins of Life*. London, U.K.: Oxford Univ. Press, 2002.
[68] M. Tooley, "Abortion and infanticide," *Philos. Public Affairs*, vol. 2, no. 2, pp. 37–65, 1972.
[69] S. Dehaene, H. Lau, and S. Kouider, "What is consciousness, and could machines have it" *Science*, vol. 358, no. 6362, pp. 486–492, 2017.
[70] A. Cleeremans, "Connecting conscious and unconscious processing," *Cogn. Sci.*, vol. 38, no. 6, pp. 1286–1315, 2014.
[71] J. J. Bryson, "Robots should be slaves," in *Close Engagements With Artificial Companions: Key Social, Psychological, Ethical and Design Issue*, Y. Wilks and J. Benjamins, Eds. 2010, pp. 63–74.
[72] T. Harford. (Oct. 11, 2016). *Crash: How Computers are Setting us up for Disaster*. The Guardian. [Online]. Available: https://www.theguardian.com/technology/2016/oct/11/crash-how-computers-are-setting-us-up-disaster
[73] T. Harford, *Messy: How to Be Creative and Resilient in a Tidy-Minded World*. New York, NY, USA: Riverhead, 2016.
[74] J. Danaher, "The rise of the robots and the crisis of moral patiency," *AI Soc.*, pp. 1–8, 2017. [Online]. Available: https://doi.org/10.1007/s00146-017-0773-9
[75] P. J. G. Lisboa, "Interpretability in machine learning—Principles and practice," in *Fuzzy Logic and Applications*. WILF(Lecture Notes in Computer Science), vol. 8256, F. Masulli, G. Pasi, and R. Yager, Eds. Cham, Switzerland: Springer, 2013, pp. 15–21.

ABOUT THE AUTHORS

Stephen Cave received the Ph.D. degree in philosophy from the University of Cambridge, Cambridge, U.K.

He was the Executive Director of the Leverhulme Centre for the Future of Intelligence, Senior Research Associate in the Faculty of Philosophy, and Fellow of Hughes Hall, all at the University of Cambridge. He then joined the British Foreign Office, where he served as a Policy Advisor and Diplomat. He has subsequently written and spoken on a wide range of philosophical and scientific subjects, including in the *New York Times*, *The Atlantic*, and on television and radio around the world. His research interests currently focus on the nature, portrayal, and governance of AI.

Rune Nyrup received the B.A. degree in philosophy from the University of Copenhagen, Copenhagen, Denmark, in 2010, the M.Phil. degree in history and philosophy of science from the University of Cambridge, Cambridge, U.K., in 2013, the MA degree in philosophy from the University of Copenhagen, in 2014, and the Ph.D. degree in philosophy from Durham University, Durham, U.K., in 2017.

Since 2017 he has been a Postdoctoral Research Associate at the Leverhulme Centre for the Future of Intelligence, University of Cambridge, U.K. His research focuses on the ethics of artificial intelligence analyzing the conceptual and philosophical foundations of debates about bias, fairness, transparency, and explainability in automated decision making systems.

Dr. Nyrup is a Member of the British Society for the Philosophy of Science. He won the prize for best Graduate Student Essay at the 2015 meeting of the European Philosophy of Science Association.

Karina Vold received the B.A. degree in philosophy and political science from the University of Toronto, Toronto, ON, Canada, and the Ph.D. degree in philosophy from McGill University, Montreal, QC, Canada, in 2017.

Since 2017, she been a Postdoctoral Research Associate with the Leverhulme Centre for the Future of Intelligence and a Research Fellow at the Faculty of Philosophy, University of Cambridge. Prior to arriving at Cambridge, she was a Visiting Fellow at Duke University, the University of Oslo, and Ruhr University, as well as a Lecturer at Carleton University. Her research interests include the philosophy of mind, consciousness, intelligence augmentation, neuroscience, neuroethics, artificial intelligence, and ethical questions related to AI.

Dr. Vold has been awarded a research grant from Duke University, a Doctoral Award from the Social Sciences and Humanities Research Council of Canada, and a Fellowship from the Wolfe Chair for Science and Technology Literacy at McGill University. She is currently a CanadaUK Fellow for Innovation and Entrepreneurship and a Member of both the Canadian and American Philosophical Associations.

Adrian Weller received the undergraduate degree in mathematics from Trinity College, Cambridge, U.K., and the Ph.D. degree in computer science from Columbia University, New York, NY, USA.

Previously, he held senior roles in finance. Currently, he is the Programme Director for Artificial Intelligence (AI) at The Alan Turing Institute, the National Institute for Data Science and AI, where he is also a Turing Fellow leading a group on fairness, transparency, and privacy. He is a Senior Research Fellow in Machine Learning at the University of Cambridge, and at the Leverhulme Centre for the Future of Intelligence (CFI) where he leads a project on trust and transparency. He is very interested in all aspects of AI, its commercial applications and how it may be used to benefit society. He advises several companies and charities.