AI & SOCIETY (2021) 36:263–276
https://doi.org/10.1007/s00146-020-01010-1

OPEN FORUM

From machine ethics to computational ethics

Samuel T. Segun[1]

Received: 2 March 2018 / Accepted: 12 June 2020 / Published online: 29 June 2020

Abstract

Research into the ethics of artificial intelligence is often categorized into two subareas—robot ethics and machine ethics. Many of the definitions and classifications of the subject matter of these subfields, as found in the literature, are conflated, which I seek to rectify. In this essay, I infer that using the term 'machine ethics' is too broad and glosses over issues that the term computational ethics best describes. I show that the subject of inquiry of computational ethics is of great value and indeed is an important frontier in developing ethical artificial intelligence systems (AIS). I also show that computational is a distinct, often neglected field in the ethics of AI. In contrast to much of the literature, I argue that the appellation 'machine ethics' does not sufficiently capture the entire project of embedding ethics into AI/S, and hence the need for computational ethics. This essay is unique for two reasons; first, it offers a philosophical analysis of the subject of computational ethics that is not found in the literature. Second, it offers a finely grained analysis that shows the thematic distinction among robot ethics, machine ethics and computational ethics.

Keywords Ethics of AI · Machine ethics · Robot ethics · Computational ethics · Autonomous intelligent systems · Artificial intelligence

1 Introduction

What really is computational ethics? In this essay, I not only answer the question of what computational ethics involves but also show explicitly why it constitutes an important frontier in the development of artificial intelligence systems (AIS) that are sensitive to human values. Unlike what is frequently found in the literature, I argue that the tag 'computational ethics' offers a plausible description of the project of embedding ethics into artificial intelligence systems, one that other commonly used appellations such as 'machine morality', 'friendly AI' and especially 'machine ethics' are too broad to convey. I offer justifications for this claim, showing how moral philosophers and indeed AI ethicists can approach and further develop computational ethics in praxis and as a transdisciplinary project.

As the hype around artificial intelligence increases and more AIS are built, the ethical burden these systems bring become apparent. Discourses around the ethical and moral implications of artificial intelligence are usually addressed under the auspices of the ethics of artificial intelligence (Russell et al. 2015). Among the earliest works that motivated the formation of the ethics of AI, Isaac Asimov's works stand out. Asimov, in his fictional texts, showed us that developing intelligent robots without some form of moral code could be catastrophic (Clarke 1993, 1994). Perhaps, his most notable contribution to the discipline was the formulation of the three laws of robotics[1] (Asimov 1950), which have since then motivated discussions around the possibility of having intelligent systems that are responsive to human ethics. Analyzing Asimov's three laws of robotics show the problems of formulating guidelines for building ethical autonomous intelligent systems.[2]

Samuel T. Segun
ssegun@uj.ac.za

[1] Department of Philosophy, University of Johannesburg, P.O. Box 524, Johannesburg 2006, South Africa

[1] As is well known, Asimov's three laws of robotics are as follows: 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law. 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. The fourth law, which is also referred to as the zeroth law states that a robot may not harm humanity, or, by inaction, allow humanity to come to harm.

[2] Clarke identifies certain constraints to Asimov's laws of robotics, which would make it computationally difficult to implement. These are the ambiguity and cultural dependence of terms used in the formulation of the laws; the role of judgment in decision-making, which would be quite tricky to implement given the degree of programming required in decision-making; the sheer complexity, this also bothers

Generally, the ethics of AI focuses on the socio-economic[3] (Smith and Anderson 2014) and legal impact[4] (Chopra and White 2011) of AI and the moral and ethical issues (Bostrom 2003) surrounding the use of these systems. Although research into the ethics of AI is often categorized into two subareas, robot ethics and machine ethics, many of the themes of these two subfields, as found in the literature, conflate. I address these subareas extensively in the next section, classifying the themes that make up these subdisciplines as distinct from each other. I in no way adduce that previous classifications of robot and machine ethics are wrong; rather, I infer and confirm that these two subareas are not as identical as they first appeared during the development phase of the discipline—ethics of AI.

My goal in this article is two-pronged. First, I intend to show the distinction between the two seemingly identical subareas of the ethics of AI. Second, I make a case for why it is reasonable to look beyond machine ethics to computational ethics. Although this may fit in as a technical piece, I take for granted that readers are aware of what some less technical terms mean, such as artificial intelligence and autonomous intelligent systems. In places where I have referred to technical terms, I offer brief definitions to put them in context.

I have structured this work into four sections. In the first, I show the thematic distinction between robot ethics and machine ethics. Second, I offer firm justifications as to why it is reasonable to now move to the more practical and concrete subfield of computational ethics. Third, I show contrasts and overlaps among these fields. Fourth, I make a case for why computational ethics is an important frontier in our pursuit to have artificial intelligence systems that are responsive and sensitive to human values and ethics.

2 From robot ethics to machine ethics

In this section, I will show the divergence between the primary concerns of robot ethics and those of the field commonly known as ‘machine ethics’. The purpose here is to lay the groundwork and show the distinction and logical relation among all subfields of the ethics of AI. As Wallach and Asaro (2017) note, “there is no firm distinction between robot ethics and machine ethics, and some scholars treat machine ethics as a subset of robot ethics” (2). It becomes expedient to carry out this type of conceptual work, giving an analytic distinction between these two oft-conflated fields. This is because as the field—ethics of AI—becomes more popular, taxonomic description of works done within these subfields become necessary to takes away the ambiguity that arises from conflating these fields.

2.1 Robot ethics

Robots, as commonly used today could mean one of the following: a mechanical device, an automated device, an electronic device, a computer program, a cybernetic device, and an artificial intelligence system. The last meaning, robots as artificial intelligence systems, is the meaning most commonly used in robot ethics. As Lin defines it, “A robot is a machine, situated in the world, that senses, thinks, and acts” (Lin et al. 2012, p. 18).

Robots, as used in robot ethics, refer to autonomous intelligent systems capable of carrying out complex actions that may have impact or consequences on humans or other morally significant beings. These systems may include humanoid assistive robots, self-driving cars, and other autonomous computer programs that interact with systems or users on the internet (Allen et al. 2000; Turkle 2006; Vallor 2011).

As an evolving subject of inquiry, robot ethics, also known as ‘roboethics’, focuses on how development in robotics research will affect ethical and social interaction. Furthermore, it addresses “what human social concerns tell us about how robots should be designed”, and how robots are to be used and treated (Veruggio and Operto 2006). These concerns are meant to influence how we design and engage with these systems.

For this section, I have mapped out three thematic areas robot ethics is most concerned with which I will discuss extensively later in the section. These are one, ensuring that the design process, creation, and purpose of artificial intelligence systems are ethical; two, issues of rights and the duties of humans toward robots and vice versa. And three, the interaction between humans and machines.

As an offshoot of engineering ethics, a major concern for robot ethicists is with developing robots that follow an acceptable set of rules (Abney 2012). Deciding on what

Footnote 2 (continued)

on having to account for all possible scenarios; the scope for dilemma and deadlock, the robot autonomy, audit of robot compliance, and scope of adaptation.

[3] Smith and Anderson in the 2014 published *Pew Research* titled “AI, Robotics, and the Future of Jobs”, discuss the economic and social impact of AI on society. As we continue to build more autonomous intelligent systems, we are likely to delegate responsibilities around security, environment, healthcare, food production etc. to these systems. These all raise concerns about the impact of AI on jobs and society.

[4] Closely linked to the moral issue with AI is the debate around its legal status, agency, and responsibility. With the imminent disruption in the transport sector by the introduction of self-driving cars, questions around who bears responsibility for harm caused by a self-driving car comes to mind. Also, there are more technical questions around insurance and liabilities that have to be addressed Chopra and White (2011).

constitutes these acceptable set of rules is subject to debate, as with many ethical issues. Nevertheless, it is important to raise these concerns.[5] It becomes appropriate to question the type of design, ethical limits, malevolent use, possible side effects, and the appropriate age to engage with these intelligence systems. As Bostrom and Yudkowsky note, the primary aim of robot ethics should be to ensure that artificial intelligence systems operate in ways that guarantee the safety of humans (2014). In this way, robot ethics can be seen, in a sense, as ethics for robotics (Boddington 2017).

As would be expected, different scholars have their perception of what they think robot ethics entails. For Peter Asaro, when we speak of robot ethics, three possible meanings are elicited (2006). First, we mean the ethical behaviors on our part that are triggered through and with the use of robots. Second, we question how best to design robots to act ethically and debate on the possibility of robots being ethical agents. Third, we consider the human–machine interaction and all its appurtenances (Asaro 2006, p. 9). I will briefly touch on the first and third classifications mentioned by Asaro in the course of this section, as they fit best into my delineation of the primary concerns of robot ethics. It is, however, expedient to state that the distinction of robot ethics from machine ethics that I have chosen places Asaro's second classification, robots as ethical agents, within the ambit of machine ethics.

Malle and Scheutz put forward two questions that, they argue, fall under robot ethics; these are ethical questions about designing, uses, and abuses of robots and questions about the moral capacities of robots (2014). For my delineation, I situate questions on uses, abuses and design of robots as concerns of robot ethics. Questions about the moral capacities of robots fit best within machine ethics.

The first concern of robot ethics is to ensure that the design, process, creation, and purpose of artificial intelligence systems are ethical. AIS must be created such that they are beneficial to humans (Floridi et al. 2018). When we create artificial intelligence systems, our aim is always to meet a need, solve a problem, or improve our efficiency and accuracy in a particular task. It becomes counter-intuitive if the ethical burden this supposed improvement brings outweighs its benefits. The implication is that creating a robot or AIS must be deliberate and well-intentioned. On the other hand, the challenge remains that we cannot properly regulate the creation of these systems yet since anyone with the requisite skills can build a robot in her garage without a licence. Hence, regulating this practice remains a hurdle to be crossed. Although we may fall short of ensuring proper regulation of the creation of AI systems, robot ethics nevertheless focuses on how best to create these systems and in turn regulate their use.

The question of design also encompasses the problem of algorithmic bias. During the process of data mining and analysis, certain unaccounted biases are embedded, which is often referred to as 'algorithmic biases' (Bozdag 2013). Algorithmic biases are implicit biases that are often unintentional yet embedded in intelligence systems, whether they are autonomous or not, in a way that affects value judgements and reinforce inequality, stereotypes, and partiality (Hajian et al. 2016; Danks and London 2017). If we consider that one of the central reasons for building artificial intelligence systems is to avoid such biases in the first place, then having algorithmic biases becomes a serious ethical issue requiring closure.

The second concern of robot ethics is with the issue of rights. By this I mean, the duties of humans toward robots. Robot ethicists have been engaged in answering the question, 'should robots have rights?' (Coeckelbergh 2010a). As absurd as this question may sound perhaps, works of literature abound of ethicists making a defence for or against robot rights in much the same way we speak of animal rights (Bryson 2010). When rights are being discussed in this sense, attention is not just on the legal rights of robots. Instead, the focus is on the moral rights robot may have or possess. This is usually because rights as a legal concept are often grounded on moral justifications. This may include, as Hohfeld (1923) says, privileges, claims, powers, and immunities. One of the reasons we are inclined to have this type of conversation is that our designs of robots have been anthropomorphised. Hence, we see them as extensions of ourselves.

However, as we edge closer to having a fully autonomous AI, the case for robot rights become more compelling. As Dashevsky (2017) points out, the European Parliament's legal affairs committee considers the idea of "electronic personhood" a substantial basis to accord advanced AI some rights. Joanna Bryson (2010) rejects this perspective, arguing that granting robots rights in some ways dehumanizes humans by emboldening poor decision-making, which is caused by the abdication of responsibilities and transferring it to robots.

In addition to this, robot ethicists are also concerned with whether the designs of artificial intelligence systems encroach on human rights. This is quite different from the first cluster of issues; the idea here is the rights of robots weighed up against the rights of humans. With the vicissitude of real-world ethical conundrums, when robots carry out their duties, there is a need to examine at the design stage when human rights and those we may accord to robots clash. The clash of rights between humans and machines is usually

[5] In much of the literature, Asmivo is unarguably seen as a forerunner in the development of guidelines to regulate the operations of autonomous intelligent systems.

hypothesized in such a way that priority is given to humans over machines.

The third concern of robot ethics is with human–machine interactions. Recent researches in robot ethics have been focused on caregiving robots for children, the disabled, elderly, and the role and morality of sex robots. The emphasis here is on the possible ethical burden this may bear on human interaction largely because activities such as caregiving and sex require some level of intimacy, delicacy, and humanness (Wallach and Asaro 2017). Should these activities be outsourced to robots? What ethical limits can be drawn concerning their deployment? With the possibility that we might have robots play a significant role in the future, such as caregiving, do they need to possess features like emotions to act morally or would they require a strictly rule-based system to act as such (Coeckelbergh 2010b)?

There are, however, some pros to having robots as caregivers. One, their judgment will not be impaired or strained by stress like human caregivers. Two, robots tend to be a hundred per cent available to carry out their duties—no bathroom breaks needed. Three, even though arguable, some care is better than no care at all. These, amongst many other reasons, show the positive value of automating care practice. However, ethicists like Sparrow and Sparrow (2006) argue that using care robots could be unethical. Their reasons are not farfetched. The dignity of the recipient of care is challenged by this deliberate outsourcing. Feelings of objectification and loneliness can be overwhelming if the human touch is elusive.

Another aspect of care robots stealthily addressed in the ethics of AI is robot companionship. By this, I refer to sex robots and other such artifacts not used in the same sense we speak of care robots for the elderly or disabled. Herein, the focus of robot ethics is with the state of relationship, value, and moral consequence interaction with sex robots may have on the recipient (Ramey 2005), as well as the adverse effects it may have on other humans (Danaher 2017; Turkle 2006). Several questions arise when engaging in this conversation, such as the range of features should affective robots possess. Can we make a morally defensible case for sex robots? How can erotic AI be created with ethical limitations to respect emotionally vulnerable individuals? Considering that humans tend to build an emotional bond with creatures outside their species, such as animals, roboticists are challenged to ensure their creations do not exploit these affective aspects of users (Sullins 2012). Hence, robots that mimic functional and affective properties of humans should be morally competent.

2.2 Machine ethics

Machine ethics, as a subarea of the ethics of AI, is often referred to as one of the following: computational ethics, machine morality, moral machines, artificial morality, safe AI, or friendly AI. The term is credited to Mitchell Waldrop (1987), who in his article titled 'A Question of Responsibility' called for a theory of machine ethics.

Machine ethics refers to ethical concerns as they relate to autonomous intelligent systems (Goodall 2014). In machine ethics, these systems are the subject of ethical debates. In other words, intelligent systems are not seen as mere artifacts but as possible new intake into our moral community and one that might have a profound impact on our legal, ethical, social, and economic landscape (Torrance 2013).

Each of the appellations for machine ethics (computational ethics, machine morality, moral machines, artificial morality, safe AI, or friendly AI) gets used differently, they are broadly construed as referring to the subfield of the ethics of AI different from robot ethics. Friendly AI is used to describe how we may build friendly artificial intelligence systems that are smarter and faster in performing cognitive tasks (Boyles and Joaquin 2019; Wallach and Asaro 2017). Safe AI, on the other hand, describes the safety of robots, especially in making decisions of moral consequence (Rodd 1995). Machine ethics is used to identify an aspect of the ethics of AI concerned with how machines might make ethical decisions (Anderson et al. 2005) and more importantly, how and when machines might be considered ethical agents (Muller 2019). Machine morality and moral machines are not used differently from machine ethics.

There are four primary themes in machine ethics. One, projects that focus on how we might program artificial intelligence systems to be ethical (Leben 2018; Arnold and Scheutz 2016). Two, projects that focus on the moral behavior or decision-making process of artificial intelligence systems (Leben 2017; Dietrich and Weisswange 2019; Van de Voort et al. 2015). Three, projects that focus on the question of artificial moral agency (Grodzinsky et al. 2008; McDermott 2008). Four, projects focused on the nature of computational consciousness and how this may influence our understanding of artificial moral agency (Gamez 2008; Chella and Manzotti 2009; Starzyk and Prasad 2011; Clowes et al. 2007). In a way, all four subprojects of machine ethics are inter-related. I will offer an analysis of each of these primary areas of machine ethics later in this section.

Machine, as used here, does not refer to just any piece of mechanical equipment used for a specific function or task; rather, it refers to a higher-order and complex equipment, gadget, artifact or algorithm that is capable of carrying out tasks independent of humans. In other words, much like those described as robots above, these are autonomous systems. Whereas the term 'machines' could be used to describe many forms of mechanical equipment used in factories and the likes, they are not to be construed as what is meant by machines when discussions of autonomous intelligent systems are carried out in the sphere of machine ethics.[6]

To appreciate the project of machine ethics, we have to consider the categories of moral agents. For Moor, these are, “ethical impact agents, implicit ethical agents, explicit ethical agents, and full ethical agents” (2006, pp. 15–18). Ethical impact agents refer to those artificial agents capable of having good or bad, praiseworthy or blameworthy consequences like the desktop printer referred to above. The ethical concerns here are on the potential impact these agents may have by their use or misuse.

Implicit ethical agents are intelligent agents designed with built-in safety, ethics, and security, such as a point of sale (POS) machine or an automated teller machine (ATM). These systems are designed in such a way that they request security codes before use, thereby protecting accounts of users and dispensing or debiting the requested amount—no more, no less. Moor argues that just as there are ethical implicit agents, there are also unethical implicit agents (2009), such as computer viruses designed for malevolent use.

In contrast, an explicit ethical agent refers to artificial intelligence systems “able to calculate the best action in ethical dilemmas using ethical principles” (Anderson and Anderson 2007, p. 15) and independent of human actors during this process. Self-driving cars and care robots can be said to fit into this category of artificial agents.

Full ethical agents possess qualities such as can be ascribed to human agents. Much like explicit ethical agents, they can not only make independent moral judgements but also often exhibit metaphysical qualities like intentionality, consciousness, free will, empathy, etc. As I write this paper, there are no known artificial full ethical agents; rather development is ongoing to develop explicit ethical agents complex enough to give us an insight on what full ethical agents might look like. Full artificial ethical agents can be said to be autonomous if they can operate independent of a “human mediator” (Chopra and White 2011, p. 2) in their decision-making process and could rightly be held accountable for their decisions.

Asaro in his article “What should we want from a robot ethic?” identifies five categories of ethical agents. These are amoral robots, robots with moral significance, robots with moral intelligence, robots with dynamic moral intelligence, and full autonomous moral agents (2006). Asaro’s categorisations are not very different from Moor’s who offers four categories of moral agents.[7]

For Floridi and Sanders (2004), we classify an artificial intelligence system as an agent depending on its level of abstraction (LoA henceforth). LoA helps to set a framework of reference when referring to a concept, subject, or definition in mathematics, logic, science, and even human interaction. In other words, “…abstraction acts as a ‘hidden parameter’ behind exact definitions, making a crucial difference” (Floridi and Sanders 2004, p. 4).

Quite obviously, systems that are explicit in design, or according to Asaro’s classification, systems with moral intelligence, all possess certain features. Floridi and Sanders argue that these features are indicative that these systems have attained certain levels of abstraction that we may consider qualifying enough for a moral agent, at least in a remote sense. These criteria/features include autonomy, interactivity, and adaptability.[8]

Systems embedded with moral intelligence or systems we consider explicit moral agents meet the first and second criteria, which are interactivity and autonomy. The basis for their moral intelligence is to autonomously make moral decisions in response to external stimuli. The third criterion, which is adaptability, is a characteristic feature of systems with dynamic moral intelligence and full moral agents. With advances in deep learning, explicit moral agents soon become agents with dynamic moral intelligence; as they interact with their environment and acquire more data, they begin to adjust their underlying moral framework.

[6] We might consider, for instance, a desktop printer as a machine but it is uniquely different from a self-driving car, which can also be said to be a machine. The difference here is the degree of autonomy of these systems and the attendant moral burden they carry. The actions of the printer may have a moral impact; an example is if it is used to print documents for whistleblowing activities. On the other hand, a self-driving car appears to carry a greater ethical burden because it is active in the moral decision-making process. As Lumbreras (2017) mentions, the goal of machine ethics is ultimately to ‘endow’ self-governing systems with ethical comportments. In the case above, a desktop printer would not count as ‘self-governing’ but a self-driving car would.

[7] Putting Moor’s alongside Asaro’s classification, amoral agents are those I have identified as ethical impact agents. Systems with moral significance are represented as implicit moral agents. Explicit moral agents are systems with dynamic moral intelligence that can make moral decisions while employing moral principles explicitly. The final type of moral agent identified by Moore is the full ethical agent, which shares human-like properties.

[8] In explicating the importance of these criteria, Floridi and Sanders note: “(a) Interactivity means that the agent and its environment (can) act upon each other… (b) Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and decoupled-ness from its environment. (c) Adaptability means that the agent’s interactions (can) change the transition rules by which it changes state. This property ensures that an agent might be viewed, at the given LoA, as learning its own mode of operation in a way, which depends critically on its experience” (Floridi and Sanders 2004, p. 7).

For artificial explicit ethical agents, their autonomy arises not only from their seeming ability to "work without human supervision" (Chopra 2010, p. 38) or to make independent moral decisions alone, but also from their ability to adapt to their environment, optimize efficiency, and make sound judgements from an array of ethical possibilities.

For now, machine ethicists are concerned with addressing ethical issues that apply to explicit or dynamic moral agents. Generally, the project of machine ethics focuses on how best we might ground ethical decisions of machines or robots on normative ethical principles. Unlike robot ethics, machine ethics focuses more on embedding ethics into autonomous intelligent systems to allow them to make ethical decisions. Below, I give an extensive analysis of the four themes of machine ethics.

First, in building AI systems that are beneficial to us, as is the concern of robot ethics, we must investigate how best to program them.[9] By programming these systems, I mean embedding an ethical principle or a moral code into these systems (Floridi and Sanders 2004; Wallach et al. 2010; Lin et al. 2012). The obvious challenge this poses is that we are inundated with several ethical theories. This is because in ethics we not only account for right or wrong actions, but also account for permissible actions. The disparities amongst ethical theories make it quite a daunting task to embed ethics or a moral code into artificial intelligence systems (Allen et al. 2000). Furthermore, ethical theories are not universally applicable, as what is true in one instance may not be in another, which implies that at least at some level, ethics is laced with subjectivity (Brundage 2014). For this reason, not many ethical theories have been applied to autonomous intelligent systems. Particular emphasis has been placed on Kantianism (Powers 2006), utilitarianism (Grau 2006; Faulhaber et al. 2019) as versions of deontological and consequentialist ethics, respectively. Quite recently, there have been attempts at developing a virtuous robot, patterned after the ethical principle of virtue ethics (Danielson 2002).

Second, machine ethicists are tasked with identifying the moral behavior or decision-making process of these intelligent systems (Moor 2006). The project of programming artificial intelligence systems to act ethically requires some features. These include what Wendell and Allen (2012) calls 'the framing problem'. Simply put, the framing problem is the challenge of teaching machines how to identify ethically significant situations. That is the "ability of AIS to recognize ethically significant situations [and] human ethical concerns into selecting safe, appropriate, and moral courses of action" (Wallach and Asaro 2017). Wallach et al. (2010) observe that accounting for how autonomous systems may factor in moral consideration into their decision-making process is what gave rise to machine ethics. As Torrance avers, "Machine ethics deals with the ways in which human-made devices might perform actions that have ethical significance" (2008, p. 495). This aspect of machine ethics touches on developing the decision-making framework of artificial intelligence systems.

Third, machine ethics is concerned with the question of artificial moral agency (Allen et al. 2000). Earlier, I had mentioned that machine ethicists do not think of machines as mere objects in discussions on AI but as subjects of this debate; Boyles (2018) argues that machine ethicists consider intelligent systems as possible moral agents. Machine ethicists are preoccupied with questioning when we might have and what possible criteria might make an artificial agent a moral agent. The obvious indication is that we do not want these systems built without some form of an ethical framework that respects human considerations (Abney 2012; Floridi and Sanders 2004; Johnson 2004; Moor 2006).

By artificial moral agency (AMA), machine ethicists are referring to the ability of a self-governing intelligent system to make moral judgements based on its notion of right and wrong (Johnson and Miller 2008). Perhaps, the major drawback with this project is with the aspect of accountability. A full ethical agent is not only able to make moral decisions but also held accountable for such decisions (Marino and Tamburrini 2006). This does not seem to be the case with artificial moral agents. This frontier, developing artificial moral agents, is important in the advancement of debates in robot ethics such as the rights of robots.

The fourth concern of machine ethics, which is closely linked with the discussions on artificial moral agency, is addressing the computational possibility of consciousness. Herein, the desire is to theorize how to build complex explicit ethical systems or full ethical systems. Although as I write this, complex explicit ethical agents do not yet exist, and neither do artificial full ethical agents, researchers are increasingly interested in building these sort of systems. One of the important frontiers to be crossed here is with accounting for consciousness (Chella and Manzotti 2009; Starzyk and Prasad 2011). Questions about machine consciousness play a significant role in the philosophy of mind. For machine ethics, it serves the purpose of addressing what might indeed constitute a moral agent and if it is at all possible to have AIS become full moral agents that can

[9] In answering the question of how to go about the embedding of ethical principles into AIS, it behoves machine ethicists to decide on the best approaches to use. So far, three approaches standout, top-down, bottom-up and hybrid. In the top-down approach, an ethical principle is selected and applied in a theoretical form to the AIS using a rule-based method such as Asimov's three laws of robots (Allen et al. 2005). The bottom-up approach, on the other hand, does not refer to any particular ethical principle; instead, through machine learning, these intelligent systems can learn subsets of ethical principles and over time integrate these into a whole and possibly unique ethical system (Wallach and Allen 2008). Then there is the hybrid approach, which simply is the fusion of the two approaches.

be accountable for their actions (McDermott 2007; Reggia 2013).

3 Why computational ethics

In this section, I make the case that using the tag 'machine ethics' is too broad and glosses over very important aspects of the ethics of AI that the term 'computational ethics' best describes. In other words, it groups specific technical areas under an otherwise distinct field. I show that the subject of inquiry of computational ethics is of great value and indeed is an important frontier in developing ethical autonomous intelligent systems. More so, I offer justifications showing that computational ethics goes beyond the theoretical limitations of machine ethics and gives a prima facie description of actualising the project of building ethical intelligence systems.

Defining computational ethics appears to be a rather easy task. Simply put, computational ethics is the project of computing ethics (Anderson et al. 2006). In other words, it is the subject of inquiry focused on actualising how artificial intelligence systems might make ethical decisions (Moor 1995). The overarching theme of computational ethics, as the name implies, is with stripping ethics of complexities and making it computable (Aaby 2005). In addition, computational ethics is concerned with the computational complexities required to build intelligent systems to make ethical decisions, as well as what might constitute the computational threshold to consider these systems as ethical artificial agents (Howard and Muntean 2016, p. 222).

Indeed, little work has been done on computational ethics and it is often conflated with machine ethics as much of its focus overlaps with machine ethics; in fact, some authors use both terms (machine ethics and computational ethics) interchangeably (Yampolskiy 2012; Allen et al. 2006). Even though much of the literature on the ethics of AI does not argue that machine ethics is distinct from computational ethics in any way, I contend that computational ethics differ slightly from machine ethics. I infer that the term 'computational ethics' possesses some technical and practical edge that machine ethics does not seem to convey. Computational ethics, which goes beyond the debates in machine ethics to its actual implementation, should play a more strategic role in the ethics of AI if we are to actualise the desire to build ethical autonomous systems.

Although usage of the term 'machine ethics' is very common in the literature, works that use 'computational ethics' are quite infrequent. Most of the authors who have opted for the use of the tag 'computational ethics' do not approach it as a less technical subject. Largely, researchers within the computer sciences, information systems, and engineering have commonly used 'computational ethics' in their attempt to explain the discipline focused on practical steps to embedding ethics into intelligent systems (Moor 1995; Lokhorst 2011). I hypothesize that the reason most AI ethicists have been unmotivated to use the appellation 'computational ethics' is that the label is perceived to rid the project of much philosophical baggage, hosting it in the domain of a practical discipline.

Unlike the sobriquet machine ethics, computational ethics put ethicists to task, ensuring that they not only discuss possible ethical theories that may be applied to AI, but also that they actively engage in the process of creating ethical algorithms or procedures. In other words, when we speak of computational ethics, we do not expect a deeply abstract endeavor riddled with philosophical jargons. The aim, rather, is to apply critical and practical models to ethical principles by maintaining logical consistency. This does not mean that we should understand computational ethics as a non-philosophical endeavour; it should rather be seen as a more practical way of applying ethics to artificial intelligence. Furthermore, it should be seen as a field resident in the intersection between ethics and other scientific disciplines focused on AI research and development such as knowledge representation and reason, computation, embodiment and logic.

It is important to identify what computational ethics is and what it is not in more detail. This is because there are possible ways the subject of computational ethics might be misconstrued. I will identify possible misinterpretations and indicate how they are distinct from the primary concerns of computational ethics.

Perhaps, the most likely field of research that may easily be conflated with computational ethics is computer ethics. Computer ethics is an aspect of applied ethics that deals with the ethical issues that surround the use of computers and other computing technologies (Forester and Morrison 1991). Some have defined it as the subject of inquiry that deals with the ethical conduct and behaviors of professionals in the computer and information technology fields (Anderson 1992; Bynum 2001). James Moor refers to it as the "analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology" (1985, p. 266). Debates around computer ethics focus on "privacy, property rights, accountability, and social value" (Johnson 2004, p. 65), cybersecurity, data usage, etc. Also, normative theories are usually applied to extant ethical issues in computer ethics, giving it a philosophical feel (van den Hoven 2010; Tavani 2002).

As shown above, computer ethics varies distinctly from computational ethics. Whereas the former deals with conducts that regulate professionals and the use of computer technologies within the field, the latter focuses on how best to create ethically aligned artificial intelligence systems.

The former is about ethical guidelines and the latter is about practical steps to codifying ethics.

Another important concern of computational ethicists is with developing the decision-making architecture of artificial intelligence systems. Even though machine ethicists study and hypothesize how AIS should make ethical decisions, the computational ethicist is tasked with developing a program to make this possible. The framework for ethical judgements and the logical and computational implications of this is put to test by the computational ethicist. To achieve this, decision-making algorithms are designed and tested; counterfactuals and 'try and except' conditions are also put to test to ensure a robust ethical system is deployed for use. This process also involves scenario planning and testing that goes beyond the armchair thought experiments that machine ethicists engage with.

As extensive work continues to go into developing decision-making algorithms for ethical AIS, the question of agency comes into play. Arguably, as many ethical theories dictate, rationality is a key factor for moral agency. In this light, computational ethicists are preoccupied with testing thresholds and conditions for that make computational rationality possible. This means that it must engage with other fields such as neuroscience, cognitive science, game theory, and economics (Lewis et al. 2014; Gershman et al. 2015). Ethics has become a recent add on because the conversation of building complex autonomous intelligent systems that are sensitive to human values has become central to the conversation.

The subject of computational rationality addresses two important aspects; these are the decision-making processes in artificial intelligence systems, as well as the development of artificial moral agents (Mabaso 2020). Looking closely, it is apparent that computational ethics fuses these two concerns. What computational ethicists do is to attempt to create systems complex enough to meet the threshold in which they may be considered as computationally rational. This could mean hardwired programming of open source robots and experimentation to see if they meet the criterion of computational rationality set by machine ethicists. This is important because rationality is an important criterion for moral agency.

Following the conversations on computational rationality is the pursuit to explore computational consciousness in robots/machines. Lokhorst (2011) notes, that the study of a robot/machine's ability to contemplate its reasoning is situated in the field of computational meta-ethics. Artificial consciousness, as it is mostly called, is a budding area in AI that has significant implications for the ethics of AI (Chella and Manzotti 2013; Cardon 2006). For this reason, computational ethicists are concerned with the possibility and experimentation of computational consciousness. They do this with the understanding that the subject of consciousness might influence our understanding of the agency and patiency of artificial intelligence systems, and how we may build them. And ultimately, this could translate to the ascription of rights and privileges in society.

My advocacy for the recognition and entrenchment of computational ethics as a subset of the ethics of AI is not for pedantic satisfaction; it is strategic for the project it sets out to fulfil, which is the development of ethical intelligence systems. This is because computational ethicists develop formal structures that can help in the implementation of the project of embedding ethics into machines. Often ignored is the fact that computational ethicists develop algorithms for decision-making that align with one or more ethical principles. As humans, we analyze all available data before making a judgement; in some instances, like the classic trolley problem, we may decide to pull the lever and let the trolley run over one person to save five. In another, we prefer not to have a hand in pushing a hypothetical 'fat man' off the bridge to save five lives. This sort of moral ambivalence is what computational ethicists contend with, seeking ways to translate abstract moral principles into computer codes.

4 Some distinctions: robot, machine, and computational ethics

A taxonomical mapping of the ethics of AI is important so we can situate research within this area appropriately and also foster collaborations across seemingly unrelated disciplines. From the above analysis of robot, machine, and computational ethics, there are obvious contrasts and overlaps as schematised in Fig. 1, which should not be ignored. In this section, I succinctly show these contrasts and where these overlaps occur, giving insights to why I believe all three interconnected subareas play important roles in the ethics of AI.

As discussed in the previous section, the thematic focus of robot ethics is threefold. First, it focuses on ensuring that the design, creation, and purpose of artificial intelligence systems are ethical. In addressing this issue, questions about autonomous weapons systems (AWS) come to mind with some ethicists suggesting that AWS should be banned entirely on grounds of their purpose not being ethical (Sauer 2016; Asaro 2012). Second, it focuses on how rights may conflict if we include AIS to our moral circle. Herein, the goal is to show justifications to accept or reject the proposition that robots should be entitled to moral rights. The third focus is on the impact of human–machine interactions. As we become heavily dependent on AIS and automate several aspects of our lives, we are faced with unique sets of issues that challenge our ethical convictions and paradigm.

Herein, questions about the moral consequences of care and sex robots on humans seem to be the top issues.

On the other hand, the thematic concern of machine ethics is fourfold. First, projects that focus on how we might program artificial intelligence systems to be ethical. The focus here is on examining ethical theories that can be best applied. Second, projects that focus on the moral behavior or decision-making process of artificial intelligence systems. An example of questions asked here would include, how should a utilitarian self-driving car act in a moral conundrum? Third, projects that focus on the possibilities and improbability of artificial moral agency. Can we ever have systems we can consider moral agents in the same sense we speak of humans? Answers to this question may rest heavily on how we respond to the question of moral responsibility and accountability. Fourth, projects focused on the nature of computational consciousness and how this may influence our understanding of artificial moral agency. The engagement with the problem here is largely theoretical.

Even though the concerns of machine ethics do appear to significantly overlap with those of computational ethics, the former only sets the theoretical foundation upon which practical computational techniques are built for application. On the other hand, the latter is more concerned with testing these hypotheses. Can we have artificial intelligence systems that meet a minimum threshold for which we can consider them computationally rational? To find answers to this, we must go beyond armchair analysis to actual experimentation of these ideas.

Computational ethics also seeks to simulate consciousness in AI. With this, we can have more evidence to point us towards the possibility that artificial consciousness is possible. Asimov's I, Robot gives us a fictional look at what it would be like to have a robot that achieved consciousness. In the story, that robot was able to defy and override the programming of VIKI (Asimov 1950).

On the overlap, for one, robot ethics questions the design and purpose of artificial intelligence systems. In other words, it shows the moral justification of having ethical compliant robots. On the other hand, machine ethics focuses on how we might go about programming these systems to ground the justifications raised by robot ethics. Computational ethicists then take all these into account to build such a system.

Two, robot ethics concern with human–machine interaction and attendant ethical issues that may arise from this interaction differs from machine ethics focus, which is on the decision-making processes of artificial intelligence systems and the effect of these processes and consequent decision on the interaction between man and machine. Computational ethicists, being aware of these ethical concerns and possible suggestions, develop ethically grounded algorithms and codes to be able to execute them during simulations and tests.

Three, overlapping the debate on rights, which is central to robot ethics, machine ethics focuses on debates about what qualifies an entity as a moral agent. This is because, for the rights of robots to be grounded, they first must be seen, to a degree, as moral agents. This means that the project of computational consciousness, which is situated in the domain of machine ethics, is instructive to how we proceed with discussions on artificial personhood and artificial agency. On the other hand, we cannot talk about artificial moral agency if we are unable to build them. This is where computational ethicists come in; they develop systems that we can say are artificial moral agency.

5 Computational ethics as an important frontier

In this section, I show why computational ethics is an important frontier in the attempt at embedding ethics into increasingly autonomous intelligent systems. I make four justifications for this claim, showing how AI ethicists can better engage with the projects at the heart of computational ethics. I contend that it is an important frontier primarily because without computational ethics we will only have theoretical discussions about building ethical AI. With computational ethics, we actually engage in this activity, programming these systems and addressing the shortcomings we confront upon their deployment.

Having shown how computational ethics differs from machine and robot ethics, it is important to note that computational ethics is not just a practical dimension of machine ethics; it is also a transdisciplinary approach to developing intelligence systems to act ethically. A computational ethics approach, unlike machine ethics, is to lay technical groundworks required for the building of ethical compliant systems. For example, in developing a self-driving car, creating algorithms that align with ethical principles are important. The computational ethicist would be concerned with creating the right algorithm that is both compliant with an ethical system and functionally dynamic to operate in the real world.

To ground computational ethics as an important and significant frontier in the ethics of AI, four justifications come to mind. First, unlike machine ethics and robot ethics, computational ethics boasts of its practical edge in actualising the goal of its project. It raises practical questions on the plausibility and, more importantly, the tractability of ethical principles as they apply to autonomous intelligent systems (Brundage 2014). As I have argued above, computational ethics put to task the desire to move beyond simple debates of methods to implementation.

Computational ethicists have shown significantly that 'moral codes' can be embedded in intelligent systems, even though it is a very difficult project to embark on. This is not

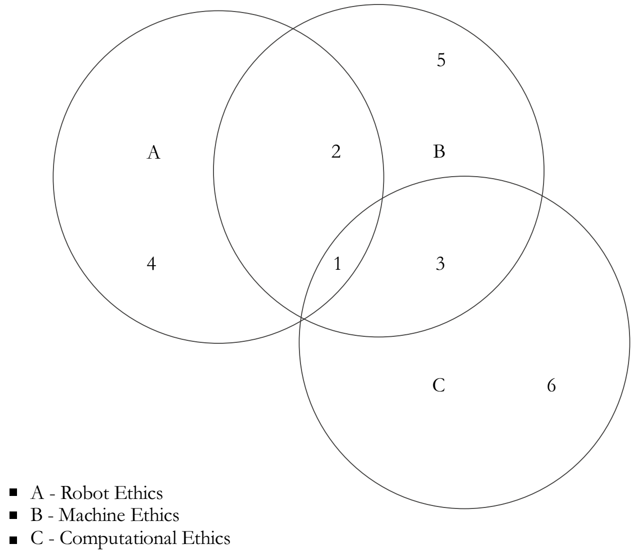

Fig. 1 This Venn diagram highlights the subareas in the ethics of AI, showing the logical relations among robot ethics, machines ethics, and computational ethics. It draws attention to the overlaps and contrast of themes in these fields. 1. Elements in A, B and C—The subject of moral agency overlaps among the three subareas. For robot ethics, it is inquiries about the rights of robots; for machine ethics, debates on what qualifies an entity as a moral agent; while computational ethics seeks to build systems that can be considered moral agents. 2. Elements in A and B—The subject of human–machine relations overlap between robot ethics and machine ethics. Robot ethics considers this from the point of safety, while machine ethics considers it from the point of ethical behaviors of AIS. 3. Elements in B and C—The subject of computational rationality and consciousness is studied by both machine ethics and computational ethics. 4. Elements in A not in B or C—The question of the ethical implications of design and purpose of creating intelligent systems is a subject of robot ethics not considered by either machine or computational ethics. 5. Elements in B not in A or C—Machine ethics studies how ethical principles can be applied to artificial intelligence systems. The focus here is to question the suitability of ethical principles. This subject is not addressed by either robot or computational ethics. 6. Elements in C not in A or B—Computational ethics, unlike robot or machine ethics, is heavily practical and seeks to program and develop robots to be ethical. It takes the theoretical work done by robot & machine ethicist into account in the building of ethical AIS

some theoretical attempt at discussing the impact of human interaction with intelligent systems but rather a pursuit to actualise the project of making ethical artificial intelligence systems. By actualising its project, I mean it queries which ethical principle might stand out as computable; in fact, computational ethicists have demonstrated this practical edge by collaborating with computer scientists and engineers to develop AIS that are responsive and, at the very least, sensitive to ethical theories in their interaction with humans and other morally significant beings (Anderson and Anderson 2007).

Second, computational ethicists have shown optimism in their work to have ethically sensitive self-driving cars and care robots. In their attempt to do this, they address some of the questions not typically asked by robot ethicists such as how might we make ethics tractable (Brundage 2014)? Can we address the framing problem, in other words, can we design intelligent systems that can identify morally significant situations (Wallach and Allen 2012)? How do we address the problem of moral uncertainty and probability (Shachter et al. 2017)? Can we have a moral justification of an action irrespective of its unintended consequences? To fully address these questions, computational ethics has to be experimental and procedural. It not only suggests ways in which answers may be given to these questions but also it models these answers in forms that are computable and applicable (Loukides 2017). Simulation modeling is one way this can be carried out, which simply, for a theorist, is a transition from thought experiments to creating these experiments with computerized logical and mathematical tools (Chung 2003).

Third, in cases where it is largely evident that an ethical principle is not calculable or possess non-procedural features, computational ethicists are tasked with designing an analytic framework to validate the usefulness of these principles. In other words, tapping into the resource of knowledge representation and reasoning (KRR), computational ethicists develop the semantic and syntactic functions required to represent these abstract ethical principles in forms that are computable for artificial intelligence systems (Levesque 1986). For example, if we have an ethical principle that dictates that "right acts are acts that cause the least harm", the computational ethicists must be able to deconstruct and unpack the concept of 'least harm', stating what the AI should consider 'least harm' in all ramification.

Fourth, computational ethics is a transdisciplinary and interdisciplinary project. Aspects of computational ethics cross disciplinary boundaries. And this is important if we have to develop ethical AIS. In Fig. 2 below, I show, using a schematic diagram, how computational ethics interacts with other disciplines in the study of AI. Even though ethics serves as the core of the subject of computational ethics, it is heavily dependent on logic to attain a formal structure. On the other hand, representing ethics within appropriate and computable syntax and semantics would require the tools used in knowledge representation and reasoning (KRR) to unpack the meanings of abstract words and concept (Baral and Gelfond 1994). The possibility of having a transdisciplinary relationship among ethics, logic, KRR makes the project of computing ethics a little more tractable than it is in abstract form.

Furthermore, in the attempt to build ethically responsible AI that fall into the category some would call artificial moral

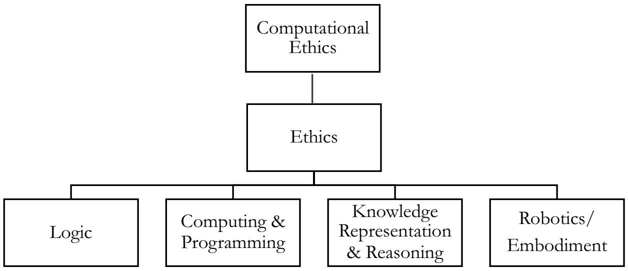

Fig. 2 This shows the interaction of key disciplines in the formation of computational ethics. Although ethics stands as the core subject of computational ethics, from the diagram above, its formalization is dependent on logic, computing, and knowledge representation and reasoning. Further actualisation is dependent on robotics/embodiment

agents, it would be important to have these agents embody a form. This is why embodiment, often grouped under robotics, is an important part of computational ethics (Parthmore and Whitby 2014).[10]

Arriving at the kind of conception of computational ethics I have suggested above would greatly improve the quest to build ethically sensitive artificial intelligence; in fact, it allows taking an experimental approach to ethical theories that would prove efficient in testing its usefulness when applied to AI. A proclivity toward this thinking prompted Ruvinsky to see computational ethics as,

> …the integration of computer simulation and ethics theory. More specifically, computational ethics is an agent-based simulation mechanism that takes a computational perspective to ethics theory. This approach uses computer modelling and multiagent systems to generate societies of agents capable of adopting various ethical principles. The principle adopted by an agent will dictate its moral action in response to a moral dilemma. By simulating the agents' application of ethical principles to moral dilemmas and observing the resulting moral landscape of a group of affected agents, we are better able to understand the social consequences of individual ethical actions' (2007, p. 1).

Ruvinsky's description offers insight into the ultimate evolution of computational ethics. In essence, the project moves beyond simply suggesting ethical principles to govern the operations of artificial intelligence systems and requires an in-depth understanding of computing. Ultimately, AI ethicists engaged with computational ethics would have to adopt procedural techniques required to compute commands into intelligent systems. A fair knowledge of machine learning becomes important in designing and applying appropriate ethical frameworks to intelligent systems.

One major objection to my proposal is that the belief that computational ethics, as I have identified it, is no different from what computer scientists and robot ethicists are currently doing. It is not that simplistic; computational ethics, as I have described, requires a rich knowledge of ethics, one that perhaps only a moral philosopher can share. With a grounding in ethics, computational ethicists would require some understanding of knowledge representation and reasoning (KRR), which could further strengthen the ethical frameworks AIS are built on.

AI ethicists ought not to see computational ethics the same way they see machine ethics. The difference bears on the practicality of the former over the later. The obvious implication of this is that we can access better insight into what is tenable and what is not if we put our hypothesis to test. AI ethicists are to understand robot and machine ethics as the first and second frontier in discussions on developing ethically compliant AIS; computational ethics then becomes the critical missing piece we need.

6 Conclusion

In this essay, I have provided an extensive and analytic distinction among the subareas of the ethics of AI—robot, machine, and computational ethics. More so, I have shown contrasts and overlaps among these disciplines, highlighting the key roles they play in achieving the goal of the discipline—ethics of AI.

Discussions about the moral justification of creating intelligent systems, the socio-ethical and socio-economic impact that human–machine interaction may have on society, the possible conflict of human and robot rights, and debates on the moral status of artificial intelligence systems, I have argued are important issues in the ethics of AI and particularly robot and machine ethics. On the other hand, attempts at programming ethics into AI systems, building an artificial moral agent, simulating consciousness in machines fall under computational ethics.

In this essay, I have offered a four-pronged justification as to why the 'computational ethics' presents a prima facie description of the project of embedding ethics in artificial intelligence systems. I have also argued that computational ethics should be seen as a separate subfield of inquiry but as an 'important frontier' in the ethics of AI. As a uniquely practical way of applying ethics to AI, computational ethics fuses relevant disciplines like computing, KRR, ethics, logic and embodiment to actualise its goal of building ethical intelligent systems. Doing computational ethics as

[10] Parthmore and Whitby make a strong case for why embodiment constitutes an important aspect in the project to build artificial moral agents. This is because embodiment appeals to the human tendency to relate and nurture, and does so regardless of the form these systems come in—biological or synthetic. Usually, we tend to care for things we anthropomorphise.

I have nudged would mean that moral philosophers would become inclined to engage ethics in quantificational forms. So far, in the literature, many ethicists are disinclined to approach the ethics of AI this way; I have shown that it remains an expedient exercise in computational ethics.

Lastly, there are two ways to consider my recommendations on the import of computational ethics to the ethics of AI. One way is to embrace it as an instructive piece and encourage collaboration among ethicists, roboticists, and computer scientists. The other is to reject it, which would imply that we continue to work in silos, each to her/his own. I prescribe the first option, as moral philosophers are often better equipped at understanding the normative aspects of ethical theories. This level of appreciation of ethics allows moral philosophers to be experts at laying the theoretic foundation upon which the computer scientists and roboticists can begin experimentation. At the same time, it affords computer scientists meaningful insight into ethics in the bid to build safe AI.

Acknowledgements I would like to thank Prof Thaddeus Metz and Prof Emma Ruttkamp-Bloem, who both took the time to read the initial drafts of this paper and made significant comments and suggestions. I would also like to thank my research group members at the Centre for Artificial Intelligence Research, University of Pretoria.

Funding None.

Compliance with ethical standards

Conflict of interest The author declares no conflict of interest.

References

Aaby AA (2005) Computational ethics. Creative commons attribution license. https://pdfs.semanticscholar.org/2db4/e8051cbbab4b916520d9ff15ef68a315a21b.pdf. Accessed 25 Sept 2019.

Abney K (2012) Robotics, ethical theory, and metaethics: a guide for the perplexed. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 35–52

Allen C, Wallach W (2012) Moral machines: contradiction in terms or abdication of human responsibility. In: Lin P, Abney K, Bekey GA (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 55–68

Allen C, Varner G, Zinser J (2000) Prolegomena to any future artificial moral agent. J Exp Theor Artif Intell 12(3):251–261

Allen C, Smit I, Wallach W (2005) Artificial morality: top-down, bottom-up, and hybrid approaches. Ethics Inf Technol 7(3):149–155

Allen C, Wallach W, Smit I (2006) Why machine ethics? IEEE Intell Syst 21(4):12–17

Anderson RE (1992) Social impacts of computing: codes of professional ethics. Soc Sci Comput Rev 10(4):453–469

Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–15

Anderson M, Anderson S, Armen C (2005) Towards machine ethics: implementing two action-based ethical theories. In Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics, (pp. 1–7).

Anderson M, Anderson SL, Armen C (2006) An approach to computing ethics. IEEE Intell Syst 21(4):56–63

Arnold T, Scheutz M (2016) Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. Ethics Inf Technol 18(2):103–115

Asaro PM (2006) What should we want from a robot ethic? Int Rev Inf Ethics 6(12):9–16

Asaro P (2012) On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. Int Rev Red Cross 94(886):687–709

Asimov I (1950) Runaround. I, robot. Bantam Dell, New York

Baral C, Gelfond M (1994) Logic programming and knowledge representation. J Logic Program 19:73–148

Boddington P (2017) Towards a code of ethics for artificial intelligence. Springer, Cham

Borenstein J, Pearson Y (2010) Robot caregivers: harbingers of expanded freedom for all? Ethics Inf Technol 12(3):277–288

Bostrom N (2003) Ethical issues in advanced artificial intelligence. Sci Fiction Philos Time Travel Superintell 2003:277–284

Bostrom N (2016) Ethical issues in advanced artificial intelligence. In: Schneider S (ed) Science fiction and philosophy: from time travel to superintelligence. Wiley, Oxford, pp 277–284

Bostrom N, Yudkowsky E (2014) The ethics of artificial intelligence. In: Frankish K, Ramsey WM (eds) The Cambridge handbook of artificial intelligence. Cambridge University Press, Cambridge, pp 316–334

Boyles RJM (2018) A case for machine ethics in modelling human-level intelligent agents. Kritike: Online J Philos 12(1): 182–200.

Boyles RJM, Joaquin JJ (2019) Why friendly AIs won't be that friendly: a friendly reply to Muehlhauser and Bostrom. AI Soc. https://doi.org/10.1007/s00146-019-00903-0

Bozdag E (2013) Bias in algorithmic filtering and personalization. Ethics Inf Technol 15(3):209–227

Brundage M (2014) Limitations and risks of machine ethics. J Exp Theor Artif Intell 26(3):355–372

Bryson JJ (2010) Robots should be slaves. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issues. John Benjamins Publishing Company, Amsterdam, pp 63–74

Bynum TW (2001) Computer ethics: its birth and its future. Ethics Inf Technol 3(2):109–112

Cardon A (2006) Artificial consciousness, artificial emotions, and autonomous robots. Cogn Process 7(4):245–267

Chan D (2017) The AI that has nothing to learn from humans. The Atlantic. https://www.theatlantic.com/technology/archive/2017/10/alphago-zero-the-ai-that-taught-itself-go/543450/. Accessed 25 Sept 2019.

Chella A, Manzotti R (2009) Machine consciousness: a manifesto for robotics. Int J Mach Conscious 1(01):33–51

Chella A, Manzotti R (2013) Artificial consciousness. Imprints Academics: Exter, UK

Chopra S (2010) Rights for autonomous artificial agents? Commun ACM 53(8):38–40

Chopra S, White LF (2011) A legal theory for autonomous artificial agents. University of Michigan Press, Michigan

Chung CA (ed) (2003) Simulation modelling handbook: a practical approach. CRC Press, London

Clarke R (1993) Asimov's laws of robotics: implications for information technology. Part 1. Computer 26(12):53–61

Clarke R (1994) Asimov's laws of robotics: implications for information technology. Part 2. Computer 27(1):57–66

Clowes R, Torrance S, Chrisley R (2007) Machine consciousness. J Conscious Stud 14(7):7–14

Coeckelbergh M (2010a) Moral appearances: emotions, robots, and human morality. Ethics Inf Technol 12(3):235–241

Coeckelbergh M (2010b) Robot rights? Towards a social-relational justification of moral consideration. Ethics Inf Technol 12(3):209–221

Danaher J (2017) The symbolic-consequences argument in the sex robot debate. In: Danaher J, McArthur N (eds) Robot sex: social and ethical implications. MIT Press, Cambridge

Danielson P (2002) Artificial morality: virtuous robots for virtual games. Routledge, London

Danks D, London AJ (2017) Algorithmic bias in autonomous systems. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 4691–4697). AAAI Press

Dashevsky E (2017) Do robots and AI deserve rights? Pc magazine. https://www.pcmag.com/article/351719/do-robots-and-ai-deserve-rights. Accessed 25 Sept 2019.

Dietrich M, Weisswange TH (2019) Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios. Ethics Inf Technol. https://doi.org/10.1007/s10676-019-09504-3

Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, Sütfeld LR, König P (2019) Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for autonomous driving vehicles. Sci Eng Ethics 25(2):399–418

Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14(3):349–379

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Schafer B (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707

Forester T, Morrison P (1991) Computer ethics: cautionary tales and ethical dilemmas in computing. Harvard J Law Technol 4(2):299–305

Gamez D (2008) Progress in machine consciousness. Conscious Cogn 17(3):887–910

Gershman SJ, Horvitz EJ, Tenenbaum JB (2015) Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. Science 349(6245):273–278

Goodall NJ (2014) Machine ethics and automated vehicles. In: Meyer G, Beiker S (eds) Road vehicle automation. Springer, Cham, pp 93–102

Grau C (2006) There is no "I" in "robot": robots and utilitarianism. IEEE Intell Syst 21(4):52–55

Grodzinsky FS, Miller KW, Wolf MJ (2008) The ethics of designing artificial agents. Ethics Inf Technol 10(2–3):115–121

Hajian S, Bonchi F, Castillo C (2016) Algorithmic bias: from discrimination discovery to fairness-aware data mining. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2125–2126). ACM.

Hohfeld WN (1923) Fundamental legal conceptions as applied in judicial reasoning: and other legal essays. Yale University Press, New Haven

Howard D, Muntean I (2016) A minimalist model of the artificial autonomous moral agent (AAMA). In 2016 AAAI Spring Symposium Series.

Johnson DG (2004) Computer ethics. In: Floridi L (ed) The Blackwell guide to the philosophy of computing and information. Wiley, Oxford, pp 65–75

Johnson DG, Miller KW (2008) Un-making artificial moral agents. Ethics Inf Technol 10(2–3):123–133

Leben D (2017) A Rawlsian algorithm for autonomous vehicles. Ethics Inf Technol 19(2):107–115

Leben D (2018) Ethics for robots: how to design a moral algorithm. Routledge, Abingdon

Levesque HJ (1986) Knowledge representation and reasoning. Ann Rev Comput Sci 1(1):255–287

Lewis RL, Howes A, Singh S (2014) Computational rationality: Linking mechanism and behavior through bounded utility maximization. Topics Cognit Sci 6(2):279–311

Lin P, Abney K, Bekey GA (2012) The ethical and social implications of robotics. MIT Press, Cambridge

Lokhorst GJC (2011) Computational meta-ethics. Minds Mach 21(2):261–274

Loukides M (2017) On computational ethics. O'Reilly. https://www.oreilly.com/radar/on-computational-ethics/. Accessed 25 Sept 2019.

Lumbreras S (2017) The limits of machine ethics. Religions 8(5) http://doi:10.3390/rel8050100.

Mabaso BA (2020) Computationally rational agents can be moral agents. Ethics Inf Technol 24:1–9

Malle BF, Scheutz M (2014) Moral competence in social robots. In Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology (p. 8), IEEE Press, Piscataway

Marino D, Tamburrini G (2006) Learning robots and human responsibility. Int Rev Inf Ethics 6(12):46–51

McDermott D (2007) Artificial intelligence and consciousness. In: Zelazo PD, Moscovitch M, Thompson E (eds) The Cambridge handbook of consciousness. Cambridge University Press, Cambridge, pp 117–150

McDermott D (2008) Why ethics is a high hurdle for AI. In North American conference on computing and philosophy. Bloomington: https://cs-www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf

Moor JH (1985) What is computer ethics? Metaphilosophy 16(4):266–275

Moor JH (1995) Is ethics computable? Metaphilosophy 26(1/2):1–21

Moor JH (2006) The nature, importance and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21

Moor J (2009) Four kinds of ethical robots. Philosophy Now 72:12–14

Müller VC (2019) Ethics of artificial intelligence and robotics. In Edward N. Zalta (ed.), Stanford Encyclopaedia of Philosophy. https://philarchive.org/archive/MLLEOA-4. Accessed 22 Sept 2019.

Parthemore J, Whitby B (2014) Moral agency, moral responsibility, and artifacts: what existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. Int J Mach Conscious 6(02):141–161

Powers TM (2006) Prospects for a Kantian machine. IEEE Intell Syst 21(4):46–51

Ramey CH (2005) 'For the sake of others': The 'personal' ethics of human-android interaction. Cognitive Science Society, Stresa, pp 137–148

Reggia JA (2013) The rise of machine consciousness: Studying consciousness with computational models. Neural Networks 44:112–131

Rodd MG (1995) Safe AI—is this possible? Eng Appl Artif Intell 8(3):243–250

Russell S, Hauert S, Altman R, Veloso M (2015) Ethics of artificial intelligence. Nature 521(7553):415–416

Ruvinsky AI (2007) Computational ethics. In: Quigley M (ed) Encyclopaedia of information ethics and security. IGI Global, Hershey, pp 76–82

Sauer F (2016) Stopping'Killer Robots': why now is the time to ban autonomous weapons systems. Arms Control Today 46(8):8–13

Shachter RD, Kanal LN, Henrion M, Lemmer JF (eds) (2017) Uncertainty in artificial intelligence 5 (Vol. 10). Elsevier, Amsterdam

Smith A, Anderson J (2014) AI, Robotics, and the future of jobs. Pew Research Center, p 6.

Sparrow R, Sparrow L (2006) In the hands of machines? The future of aged care. Mind Mach 16:141–161

Starzyk JA, Prasad DK (2011) A computational model of machine consciousness. Int J Mach Conscious 3(02):255–281

Sullins JP (2012) Robots, love, and sex: the ethics of building a love machine. IEEE Trans Affect Comput 3(4):398–409

Tavani HT (2002) The uniqueness debate in computer ethics: what exactly is at issue, and why does it matter? Ethics Inf Technol 4(1):37–54

Torrance S (2008) Ethics and consciousness in artificial agents. AI Soc 22(4):495–521

Torrance S (2013) Artificial agents and the expanding ethical circle. AI Soc 28(4):399–414

Turkle S (2006) A nascent robotics culture: new complicities for companionship. American Association for Artificial Intelligence Technical Report Series AAAI. https://www.aaai.org/Library/Workshops/2006/ws06-09-010.php. Accessed 22 Sept 2019.

Vallor S (2011) Carebots and caregivers: sustaining the ethical ideal of care in the twenty-first century. Philos Technol 24(3):251

Van de Voort M, Pieters W, Consoli L (2015) Refining the ethics of computer-made decisions: a classification of moral mediation by ubiquitous machines. Ethics Inf Technol 17(1):41–56

Van den Hoven J (2010) The use of normative theories in computer ethics. In: Floridi L (ed) The Cambridge handbook of information and computer ethics. Cambridge University Press, Cambridge, pp 59–76

Veruggio G, Operto F (2006) Roboethics: a bottom-up interdisciplinary discourse in the field of applied ethics in robotics. Int Rev Inf Ethics 6(12):2–8

Waldrop MM (1987) A question of responsibility. AI Mag 8(1):28–28

Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford

Wallach W, Allen C (2012) Hard problems: framing the Chinese room in which a robot takes a moral Turing test. https://wendellwallach.com/wordpress/wp-content/uploads/2013/10/Hard-Problems-AISB-IACAP2012-Wallach-and-Allen.pdf. Accessed 25 Sept 2019.

Wallach W, Asaro P (2017) Machine ethics and robot ethics. Routledge, New York

Wallach W, Franklin S, Allen C (2010) A conceptual and computational model of moral decision making in human and artificial agents. Topics Cognit Sci 2(3):454–485

Yampolskiy RV (2012) Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In: Müller VC (ed) Philosophy and theory of artificial intelligence. Springer, Berlin, pp 389–396

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