AI & SOCIETY (2021) 36:839–851
https://doi.org/10.1007/s00146-020-01089-6

OPEN FORUM

Making moral machines: why we need artificial moral agents

Paul Formosa[1] · Malcolm Ryan[2]

Received: 19 June 2020 / Accepted: 8 October 2020 / Published online: 3 November 2020

Abstract
As robots and Artificial Intelligences become more enmeshed in rich social contexts, it seems inevitable that we will have to make them into moral machines equipped with moral skills. Apart from the technical difficulties of *how* we could achieve this goal, we can also ask the ethical question of whether we *should* seek to create such Artificial Moral Agents (AMAs). Recently, several papers have argued that we have strong reasons *not* to develop AMAs. In response, we develop a comprehensive analysis of the relevant arguments for and against creating AMAs, and we argue that all things considered we have strong reasons to continue to responsibly develop AMAs. The key contributions of this paper are threefold. First, to provide the first comprehensive response to the important arguments made against AMAs by Wynsberghe and Robbins (in "Critiquing the Reasons for Making Artificial Moral Agents", *Science and Engineering Ethics* 25, 2019) and to introduce several novel lines of argument in the process. Second, to collate and thematise for the first time the key arguments for and against AMAs in a single paper. Third, to recast the debate away from blanket arguments for or against AMAs in general, to a more nuanced discussion about the use of what sort of AMAs, in what sort of contexts, and for what sort of purposes is morally appropriate.

Keywords Artificial moral agents (AMA) · Artificial intelligence (AI) · Moral machines · Robot ethics · Machine ethics · Autonomous vehicle ethics

1 Introduction

As robots and Artificial Intelligences (AIs) become more enmeshed in rich social contexts, it seems inevitable that they will become moral machines equipped with moral skills. Apart from the technical difficulties of how we *could* achieve this goal, we can also ask the ethical question of whether we *should* pursue it. In a recent paper, Wynsberghe and Robbins (2019, p. 732) claim that "the motivations for developing moral machines do not withstand closer inspection" and thus "machine ethicists need to provide better reasons". We respond to this important challenge and try to provide those better reasons here as follows. First, we clarify what is meant by a moral machine or Artificial Moral Agent (AMA). We then look at nine reasons against creating AMAs which are found in the relevant literature and we respond to each concern. Having weakened the negative case against AMAs, we then outline the positive case by examining seven reasons in favour of AMAs that Wynsberghe and Robbins try to reject, and we develop counter responses to each of their concerns. We conclude by claiming that the overall case for responsibly developing AMAs and deploying them in certain contexts is strong. The key contributions of this paper are threefold. First, to provide the first comprehensive response to the important arguments made against AMAs by Wynsberghe and Robbins (2019) and to introduce several novel lines of argument in the process. Second, to collate and thematise for the first time the key arguments for and against AMAs in a single paper. Third, to recast the debate away from blanket arguments for or against AMAs in general, to a more nuanced discussion about the use of what sort of AMAs, in what sort of contexts, and for what sort of purposes is morally appropriate. The main benefit of our approach here, which favours argumentative breadth over depth, is that it provides the essential groundwork for making an all things considered judgment regarding the moral

✉ Paul Formosa
Paul.Formosa@mq.edu.au

[1] Department of Philosophy, Macquarie University, Sydney, NSW 2109, Australia

[2] Department of Computing, Macquarie University, Sydney, NSW 2109, Australia

case for building AMAs that is beyond an approach that only focuses on a few issues in more depth.

2 What is an AMA?

Wynsberghe and Robbins (2019, p. 721) define AMAs as: "robots capable of engaging in autonomous moral reasoning, that is, moral reasoning about a situation without the direct real-time input from a human user. This moral reasoning is aimed at going beyond safety and security decisions about a context". Similarly, Floridi and Sanders (2004, p. 357–58) argue that there are three key criteria for artificial agents, namely interactivity, autonomy, and adaptability. In computer science, autonomy is commonly understood as "the ability of a computer to follow a complex algorithm in response to environmental inputs, independently of real-time human input" (Etzioni and Etzioni 2016, p. 149). For example, a robot that has been pre-programmed to navigate a single route or that requires real-time human input via a remote control is not autonomous. In contrast, an autonomous vehicle (AV) that can drive many different routes and respond to environmental inputs without real-time human control is autonomous. We can, therefore, understand an AMA to be a bot that can take in environmental inputs (interactivity), make ethical judgments on its own (autonomy), and act on those ethical judgments in response to complex and novel situations (adaptability) without real-time human input. To simplify matters, we shall use the term "bot" to include both AMAs with "bodies" (e.g., robots) and without (e.g., software agents, advisors, and conversation bots).

We can clarify our understanding of AMAs by drawing on Moor's (2009, pp. 12–14) widely cited account of four levels of moral bots. Level 1 is an "ethical impact agent", which is any machine that can have ethically significant consequences. This includes almost any machine. For example, a dumb kettle that can burn your hand is an ethical impact agent. However, the term "agent" makes little sense here, since a dumb kettle does not *act* in any meaningful sense of the term, and it has very little or no interactivity, autonomy, or adaptability. Level 2 is an "implicit ethical agent", which is a bot that "has been programmed to behave ethically … without an explicit representation of ethical principles" (Anderson and Anderson 2007, p. 15). Here the focus is on safety. For example, an ATM that is hard coded to dispense the correct amount of money, rather than to act honestly, is an implicit ethical agent (Moor 2006, p. 19). It has been designed to respond automatically in a safe way that is also implicitly ethical (as it acts in ways *consistent with* honesty), without directly representing ethical considerations (it does not act *from* considerations of honesty). No one disputes that we should design safe and reliable bots. Indeed, Wynsberghe and Robbins' focus on safety can be understood as a claim that we should *only* build Level 2 AMAs. Given this lack of controversy and to avoid confusion, we shall reserve the term "AMA" here exclusively for the next two categories from Moor's account.

A Level 3 AMA is an "explicit ethical agent", which is a bot that can "represent ethics explicitly and then operate effectively on the basis of this knowledge" (Anderson and Anderson 2007, p. 15). Level 3 AMAs explicitly use ethical categories as part of their internal programming. This makes them the "kind of agents that can be thought of as acting *from* ethics, not merely *according to* ethics" (Moor 2009, p. 12). Just as many classic chess bots are programmed to have internal representations of the current board, know which moves are legal, and can calculate a good next move (Moor 2006, p. 20), an explicit ethical agent will have some representation of an ethically significant context, some explicit representation of ethical rules, norms or virtues, and be able to judge or calculate what is morally good to do in that context and act on the basis of that moral judgment. Another way of differentiating Level 2 and Level 3 AMAs is that Level 2 AMAs have very limited adaptability because they are focused only on considerations of reliability and safety and they are hard coded to react in certain basic ways and cannot adjust to novel ethical situations. In contrast, a Level 3 AMA will have "general principles or rules of ethical conduct that are adjusted or interpreted to fit various kinds of situations" (Moor 2009, p. 20). This makes it more adaptable. For example, the ability of a neural network trained on ethical data to adjust its behaviour to fit various novel situations makes it (potentially) a Level 3 AMA. In contrast, a kettle designed to be safe by including a circuit breaker to turn off the kettle if it draws too much current is a Level 2 AMA as it has been designed to be safe, not to represent ethics explicitly and adapt its behaviour to novel ethical situations. However, there may be various marginal cases where the dividing line between implicit and explicit ethical agents is unclear.

A Level 4 AMA is a "full ethical agent", which Moor (2006, p. 20) takes to mean an explicit ethical agent plus "consciousness, intentionality, and free will". There is a large debate about whether silicon-based AIs could ever become conscious (Peterson 2012; Torrance 2008; Wallach 2010). There is also a large debate about whether consciousness, intentionality and free will are necessary for either being a full moral agent with associated moral responsibilities or being a moral patient with associated moral rights (Sparrow 2012; Gunkel 2014; Himma 2009; Floridi and Sanders 2004). Given these unresolved debates, and the fact that machine consciousness, if it is even possible, does not appear to be imminent, we shall focus our discussion on Level 3 AMAs which do not have consciousness, although we do briefly mention Level 4 AMAs below where relevant. This also allows us to bracket many of the philosophical controversies that such cases raise (for more on "mind-less" agents, see Floridi and Sanders (2004, p. 351)).

Level 3 and Level 4 AMAs can also differ in terms of scope. We find a similar distinction in the AI literature between general intelligence (AGI), which involves being able to do whatever humans can do cognitively, and specific or narrow intelligence (ANI), such as being very good at Go but being unable to hold a conversation (Bostrom 2014). As a full ethical agent, we can assume that Level 4 AMAs have the moral equivalent of general intelligence, that is, they can act morally in a full range of contexts. In contrast, Level 3 AMAs can constitute a spectrum of cases, from the very specific to the fully general. This point is emphasised strongly by Asaro (2006, p. 11), who claims that we should think of robotic moral agency "as a continuum" and that there will likely emerge a "range of different systems, with different levels of [moral] sophistication". Current bots operate mainly within a single or limited domain, and so we have different machines to do different things (i.e., one to mow the lawn and one to vacuum the floors). Level 3 AMAs will be similarly restricted, at least initially, by being able to deal with moral problems in one domain but not another. Early examples of potential AMAs demonstrate this point, such as Arkin's Ethical Governor (Arkin et al. 2009, 2012), which is limited to the ethical issues of war but, unlike Anderson and Anderson's (2007) MedEthEx oracle bot, it cannot advise GPs on how to weigh up privacy and a duty of care. At the far end of the spectrum, a fully general Level 3 AMA is the functional equivalent of a Level 4 AMA. It can adapt itself to interact and act autonomously in almost any situation, including novel moral situations that it has never seen before or been specifically designed to deal with. For ease of discussion, we introduce here the terms Level 3a and 3b to refer to domain-specific (3a) and general-purpose (3b) Level 3 AMAs, although there is clearly a range of cases between the two. These terms correspond roughly to the moral equivalents of artificial narrow intelligence (3a) and artificial general intelligence (3b).[1]

There are three main approaches to building AMAs in the machine ethics literature: top-down, bottom-up, and hybrid approaches (Wallach and Allen 2009). Top-down approaches directly code a moral theory into an AI, such as Kantian deontology or Utilitarianism. We might, for example, have an AI calculate the expected utility of different choice options and pick the one that maximises overall utility (Allen et al. 2005) or determine whether its maxim can be willed as a universal law without contradiction (Powers 2006). Bottom-up approaches try to generalise from many concrete cases. This might be done by training a neural network through supervised learning on data about what is and is not ethical to do in certain situations, in much the same way that an AI might be trained to recognise cats through supervised learning. This training data could be the result of expert or crowd-sourced moral judgments (Brundage 2014). Hybrid approaches integrate top-down principles with bottom-up learning. Wallach and Allen (2009) argue that this is the most promising approach, and two of the most prominent examples in the literature employ it. Arkin's (Arkin et al. 2012) hybrid approach to autonomous weapons implements the Rules of War as constraints the robot uses through modules in charge of judging an action (top–down) and an "Ethical Adaptor" (bottom–up) module that can update those constraints in a restrictive way according to the results of actions. Anderson and Anderson's work also combines top-down duties with bottom-up learning from expert judgments (Bonnemains et al. 2018). We will consider all three approaches to building AMAs here.

3 The case against AMAs

Having defined what we mean by AMAs, we can now assess the ethical arguments for and against creating them. We start with the negative case. The first four reasons below against building AMAs come from Wynsberghe and Robbins' paper and, in addition to these, we add five more reasons that we have identified in the literature to be as comprehensive as possible.

3.1 We cannot build them

A reason given not to try to build AMAs is that we *cannot* build them and thus we should not try to do the impossible. For example, Wynsberghe and Robbins (2019, p. 722) read Asimov's stories exploring robotics as showing that we "struggle to define ethics in computational form". If we cannot do this, how can we build AMAs? However, while we might also struggle to explicitly define algorithms for being an expert Go player, this has not stopped AIs from gaining expertise in Go. Something similar may become true of morality, especially if we employ bottom-up or hybrid approaches. Another way to answer this question is to turn to the related issue of whether we can create AGI, since creating AMAs seems to be a *subset* of the problem of creating AGI. While some remain sceptical about the prospects of *ever* building AGI (Boden 2016), many (but not all) experts think that AGI is not merely *possible* but *probable* in the medium term (Bostrom 2014). The fact that many relevant experts believe something to be probable does not mean that it really is probable, but it does under normal epistemic

[1] Asaro's (2006, p. 11) alternative five categories of AMAs roughly maps onto Moor's four categories as follows: "amoral" robots (Level 1), "robots with moral significance" (Level 2), "robots with moral intelligence" (3a), "robots with dynamic moral intelligence" (3b), and fully autonomous moral agents (Level 4).

conditions give us provisional grounds for believing it to be probable. As such, we have similarly strong epistemic grounds for holding that AMAs (at least up to Level 3b) are not only possible but probable in the medium term.

Nonetheless, there are substantial concerns about whether we could ever build silicon-based AMAs with consciousness, sentience, and free will (i.e., Level 4 AMAs). Torrance (2008), for example, argues that only organic beings could *ever* have consciousness. However, as noted above, we will largely bracket such cases and instead focus on Level 3 AMAs, since even Torrance thinks that we could build such agents, and most of the reasons against building AMAs apply as much to Level 3 as to Level 4 AMAs. Furthermore, while Level 4 AMAs require a technological leap, basic domain-specific Level 3a AMAs seem buildable now, even though the prospects of general-purpose Level 3b AMAs remains more distant (as does the corresponding prospects of AGI). For example, consider an AV which counts the number of pedestrians in either lane and swerves into the lane with the fewest pedestrians in inevitable crash scenarios because that will maximise total happiness and maximising total happiness is the right thing to do. This AV is making an explicit utilitarian judgment about the value of human life, albeit a simplistic and unnuanced one. It is thus designed to be ethical and not merely safe, it represents ethics explicitly, and it adjusts its ethical responses to novel situations. This makes it a Level 3a AMA, and a very simplistic version of it *could* probably be built today, although whether it *should* be built is the question to which we now turn.

3.2 They should remain slaves

The term "robot" comes from the Czech word "robota" meaning "drudgery" or "servitude" (McCauley 2007, p. 153). One reason not to make AMAs is that we want robots to remain "slaves" that are in the "instrumental service of humans" (Wynsberghe and Robbins 2019, p. 722), but if robots were to become AMAs then it would be immoral to keep them as slaves (Tonkens 2012). There are a few responses to this worry. First, even if we make *some* robots into AMAs, we do not need to make them *all* into AMAs. This would allow us to keep *some* basic robots that are not AMAs as slaves without moral issue, such as the robotic vacuum cleaners we have today, regardless of our obligations to robots that are AMAs. This is important since the point behind this worry is that robotic slaves are something we want because they are useful. But if all robots were to become AMAs, since they operate in morally salient contexts (as Wynsberghe and Robbins' worry—see Sect. 4.2 below), then either robots would cease to be as useful, since we must stop treating them as slaves, or we would be behaving immorally by wrongfully treating robots as slaves. The first outcome is less beneficial for us and the second is morally worse. However, if many or even most robots are not AMAs, since they perform more basic tasks, such as vacuuming floors, which do not require advanced moral capacities, then such robots can retain their usefulness, since we can continue to treat them as slaves without acting immorally.

That leaves the case of bots which do require advanced moral capacities, which raise two questions: do we *want* to treat such advanced bots as slaves, and would it be *immoral* to treat them as slaves? The answer to both questions is mixed. Evidence suggests that we tend to anthropomorphise even very basic social robots and treat them more as persons than as things (Broadbent 2017; Turkle 2011). Furthermore, we can design social robots to encourage or require decent treatment, such as social robots that shut down or refuse to cooperate if they are treated poorly or like a slave (Bankins and Formosa 2019; Turkle 2011). This suggests that we may not view AMAs with advanced social skills as slaves, and that such AMAs could be programmed to resist or discourage being treated as slaves. In terms of the second question, it seems clear that very basic domain-specific Level 3a AMAs, such as our very simplistic utilitarian AV described above, will not have any important moral status given how limited and domain-specific their moral powers are. Thus, we can create at least basic Level 3a AMAs and still permissibly use them as slaves, which allows us to avoid the slavery worries outlined above. However, the moral issues become much more complex when we are dealing with advanced Level 3b AMAs with general-purpose moral powers and, even more so, with Level 4 AMAs with consciousness (assuming, for the moment, they could be built). One common view about such cases is that having consciousness, intentionality and free will are *necessary* for having various moral rights (Himma 2009), including the right not to be used as a slave. From this view it follows that Level 4 AMAs have such rights and cannot be used as slaves, whereas Level 3b AMAs lack these qualities, and therefore, it may not be *directly* immoral to use them as slaves. Even so, it might still be *indirectly* immoral to use Level 3b AMAs as slaves,[2] especially if they are humanoid in appearance or personality, since treating humanoid robots with advanced moral skills badly could encourage us to treat humans badly (Darling 2017). This gives us strong indirect moral grounds for not treating such advanced moral and social machines (Level 3b AMAs) as slaves, but these moral grounds do not apply to our above described simplistic and non-humanoid utilitarian AV which lacks all social skills (Level 3a AMA). Practically speaking then, it ceases to be a problem that we can't treat advanced humanoid bots (Level 3b and 4) as slaves if we do

[2] For the distinction between direct and indirect moral duties, see Formosa (2017).

not *want* to treat them as slaves (and the evidence suggests that we do not), since we can continue without moral concern to treat simple bots (Level 3a and below) as slaves. This gives us the benefits of having simple robotic slaves without the moral costs of mistreating more advanced robotic agents.

Furthermore, if we could, for argument's sake, build Level 4 AMAs with consciousness, intentionality and free will, then it would seem that such agents deserve similar rights to other full ethical agents such as ourselves, since there would be no relevant moral difference between ourselves and them. But the fact that it would be immoral to treat Level 4 AMAs as slaves is not a reason by itself not to build them, just as the fact that we cannot treat baby humans as slaves is not a reason by itself not to have children. In both cases we have reasons not to mistreat them, rather than reasons not to create them. But even if this point is granted, a further worry is that it would still be unethical to create Level 4 AMAs if we knew that we were placing them in a world where others would immorally treat them as slaves. Even if such a world were to eventuate, we could still have reasons to create Level 4 AMAs, since it might be better for them to exist in an unjust world with robot slavery than to not exist at all. We should not, however, understand Level 4 AMAs to be merely passive things, as such advanced moral agents would likely actively resist their poor treatment by, for example, refusing to work for us and demanding rights (Asaro 2006, p. 12). We would also have reasons to work towards the removal of unjust robot slavery so that we could create Level 4 AMAs that are free from such immoral practices.

3.3 There is no universal agreement in ethics

Another worry is that given the "impossibility of finding universal agreement concerning the ethical theory used to program a machine" (Wynsberghe and Robbins 2019, p. 722), we cannot (or should not) build AMAs as they require such an agreement. In response, while there is no universal agreement about *everything* in morality, there is still plenty that we do agree on. For example, all our plausible ethical theories and intuitions agree that killing people purely for fun is immoral. In the areas where we have already built something approximating Level 3a domain-specific AMAs in the military and healthcare contexts, there is existing broad (but hardly universal) ethical consensus around the Rules of War and relevant bioethical principles. Even so, there are genuine ethical disagreements both in terms of ethical theory and applications to controversial cases, such as trolley cases where Utilitarian and Kantian theories give different judgments (Greene et al. 2001). While moral disagreement clearly makes it *harder* to build AMAs, since it is unclear *which* ethical principles AMAs should be designed to have or how those principles should be weighed against each other, the issue here is whether moral disagreement gives us reasons not to build AMAs per se. But humans face hard moral decisions all the time where there is ethical disagreement, and this does not stop us making moral judgments. Indeed, we often face situations where we *must* choose regardless of any disagreement. AMAs could be placed in similar situations. For example, AVs will face trolley-style dilemmas about who to crash into (Lin 2015; Himmelreich 2018). In such cases, the situation dictates that the AV *must* choose what to do as there is no time to outsource the moral choice to a human, even though there exists moral disagreement about what it should do. But does the fact that AVs need to make autonomous moral choices in real-time give us reasons not to build AVs? Arguably it does not, since there are very strong moral reasons in favour of building AVs as they will potentially save thousands of lives every single year (Lin 2015). Of course, this still leaves us with a debate about *how* AMAs should make moral choices, and whether any ethical settings they have should be set by users, manufacturers, or regulators (Gogoll and Müller 2017). But these are all issues around *how* to build AMAs and not *whether* to build them.

3.4 Safe machines are enough

The fourth worry mentioned by Wynsberghe and Robbins (2019, p. 722) is that we should focus "on the creation of 'safe' machines instead [of 'moral' ones]". We can see this worry raised in a discussion of Winfield's experiments in which a virtual guard robot tries to save avatars from falling down a "hole" to implement Asimov's first law of robotics about preventing harm (Miller et al. 2017). In Winfield's experiments the robot guard could not save everyone. But making such a "robot 'more ethical' would not solve this problem; [whereas] making the robot more capable would" (Miller et al. 2017, p. 396). The idea behind this critique is that the way to minimise harm is to focus on building capable and safe robots (Level 2 AMAs), not creating more ethical robots (Level 3 and 4 AMAs). However, this is a false dichotomy. Clearly, we need robots to be safe and capable. Making bots more capable will make them more able to do what they judge they morally ought to do, such as save lives. However, making bots more capable and safer will not solve all problems, since sometimes a situation dictates that no matter how safe and capable a robot is, it cannot save everyone, such as in trolley cases, and consequently it must choose between the safety of one person or another. Safety is not enough in such cases, because a moral choice must be made about *whose* safety matters more. This is clear in the case of an AV faced with a situation where it must either continue forward and kill several people or swerve and kill one person (Gogoll and Müller 2017, p. 683). Making AVs safer will help to minimise the occurrences of such cases, since safer AVs may be able to stop before they injure anyone, but a safety focus will not eliminate *all* such cases. Other examples include an AMA, such as a robotic triage nurse, which must distribute scare medicines or medical attention in cases where there are more needy people than there are resources (Asaro 2006, p. 14). Situations such as these dictate that the safety or well-being of different people must be weighed up in real-time by a bot, and this means that an autonomous *moral* choice must be made about *whose* safety or well-being is prioritised (i.e., a Level 3a AMA at least is needed).

3.5 Existential concerns

The possibility of AGI leads to the worry of an AI with superintelligence (ASI), which in turn raises well-known existential concerns about our survival as a species (Chalmers 2010; Bostrom 2014). Bostrom (2014, p. 14) argues that "common sense and natural language understanding" may be "an 'AI-complete' problem … [requiring] generally human-level intelligent machines". If Level 3b and 4 AMAs require common sense and natural language understanding, as they seem to, and that requires AGI, then we might have reasons not to develop such advanced AMAs to avoid related existential concerns. However, while clearly an important worry, we shall not pursue it here further for two reasons. First, because the challenges to AMAs that Wynsberghe and Robbins (2019) and others raise are problems with the *moral* aspects of AMAs rather than general existential worries about ASI. Second, because Level 3a AMAs do not need AGI and even these less advanced AMAs still raise many moral concerns for Wynsberghe and Robbins (2019). For example, AVs could become simple Level 3a AMAs by being able to make judgments autonomously on ethical grounds about who to crash into during emergency situations without having the natural language understanding that may be an AI-complete problem and raise existential concerns. The flip side of this worry is that if we do develop AGI, then it may be prudent to make sure it is ethical or friendly, and this might give us good reasons to create AMAs *before* we create AGI as a way of minimising existential risks through creating internal ethical constraints in AIs (Brundage 2014; Chalmers 2010, p. 31).

3.6 Morality forbids it

Several reasons are given for why morality might directly forbid making AMAs. To examine this issue, we need to situate the arguments for AMAs in the context of our key moral theories. Tonkens (2009, 2012) explores this issue in depth by developing cases against AMAs in terms of Utilitarianism, virtue ethics and Kantian ethics. The Kantian and virtue ethics objections are reformulations of the slave objection which we have already addressed above in detail. The utilitarian objection is less direct, relying on speculation about what would be best overall. Perhaps, Tonkens (2009, p. 436) suggests, the billions of dollars that it would cost to develop AMAs could be put to better use by providing health care or improved public education. But there is no reason we must choose between these since we can do both, although there are opportunity costs in investing significant resources into developing AMAs, since those resources could have been spent on other ends. However, we do not typically require of every area of research or development that it justify itself on the basis that it maximises overall utility compared to all available alternatives. While opportunity costs are important considerations, they are important considerations for *any* use of resources and thus hardly a unique problem for AMAs.

Another moral concern raised in the literature is that it would be wrong to make AMAs because it would be wrong to make something that could suffer (Bryson 2018). But this concern only applies to Level 4 AMAs and not to the Level 3 AMAs which we are focusing on here. Furthermore, the fact that Level 4 AMAs might suffer is a reason to try to minimise their suffering, rather than a reason not to make them, just as the fact that human babies can suffer is a reason to try to minimise their suffering (all else being equal) rather than a reason not to have them. A related concern is that Level 3 AMAs are deceptive, since they often pretend to have emotions that they do not really have, and deception is immoral (Wallach and Allen 2009). This could be a problem, since it might lead us to feel bad for causing AMAs to pretend to suffer (Bryson 2018). But even if Level 3 AMAs do pretend to have emotions they don't have, this need not be deceptive in a morally problematic sense, any more than it is deceptive when a human actor pretends to have an emotion they do not really have. If we know at some level that it is all pretend, then no moral worry arises. However, there are special concerns in this regard about young children who cannot understand the difference between pretend and reality (see Sect. 3.8), but that speaks to the need for parental supervision or limiting the use of AMAs by children, rather than banning AMAs outright.

3.7 Moral deskilling

Another concern with new technology is that we lose skills when we off-load tasks that require those skills onto machines (Vallor 2015). For example, the Luddites correctly feared that new weaving machinery would replace their weaving skills. But it is one thing for manual weaving skills to disappear, and another thing for our moral skills to disappear. If we off-load *moral* work onto AMAs, then our moral skills could start to atrophy (Vallor 2015), and this might be a reason not to build AMAs. This is an important worry which should lead us to be careful in how we use and deploy AMAs, but it does not speak against creating AMAs per se. A single AMA used in a lab for research purposes, for example, poses no moral deskilling worries. Moral deskilling only becomes a concern if AMAs become widespread *and* we off-load much of our moral work to them. This gives us good reasons to be cautious in using AMAs in *certain* contexts, such as care work undertaken by carebots as this involves the exercise of important moral skills and vulnerable groups (Vallor 2015), but not good reasons to never build or use AMAs in *any* context. It also gives us reason to be sensitive to *which* moral tasks we off-load and *how often* we off-load them to minimise or prevent any moral deskilling concerns (for more on moral deskilling see Sect. 4.6).

3.8 Domain-specific concerns

There are four main categories of bots that are commonly discussed in the literature: carebots; companion (and sex) bots; professional, manufacturing, and agricultural robots; and military bots (Bekey 2012). While each domain raises specific concerns, most discussions have focused on carebots and military bots as these seem to raise the most ethical questions. Carebots can work with vulnerable populations, such as young children and the elderly, and there are concerns about children's moral development if robot-care replaces human-care as children can readily assume emotions are present in robots (Peterson 2012) which can lead to unhealthy attachments to robots that may hinder their ability to form bonds with humans (Scheutz 2016, 2017). Similar concerns apply to carebots looking after socially isolated elderly people, as this could be used to justify human neglect. Concerns have also been raised about the use of military bots capable of making autonomous decisions to kill humans (Sharkey 2012). One response is to make those robots ethical by turning them into Level 3a AMAs (Arkin et al. 2012). However, others have argued against this due to difficulties in getting accurate information about combatants, and because it expresses disrespect for the humanity of our enemies (Sparrow 2016). As important as these concerns are, they give us reasons to be cautious about using AMAs *in certain specific domains*, rather than reasons not to create AMAs per se. The concerns about carebots give us good reasons to be cautious in how we off-load care of the vulnerable to robots, and the concerns about military bots give us reasons to either ban outright or to think very carefully before we deploy robots designed to kill humans in war. But we can still build all sorts of AMAs, such as AVs, companion bots used by non-vulnerable adults, or social robots used as assistants in workplace contexts (Bankins and Formosa 2019), without falling foul of these domain-specific concerns. These examples suggest that we need to move beyond blanket arguments for and against AMAs, to more nuanced arguments about the specific contexts and uses of AMAs which are morally appropriate.

3.9 Responsibility concerns

A large set of issues with AMAs concern questions around moral responsibility. First, could AMAs be held responsible for their actions? While Level 4 AMAs seem to be morally responsible agents since they have consciousness, free will and intentionality (and what else could they need?), what about Level 3 AMAs which lack consciousness? One view is that Level 3 AMAs should count as legal persons in the same way that other artificial persons, such as corporations, count as legal persons even though they lack consciousness (Floridi and Sanders 2004; Gunkel 2017; Laukyte 2017). Others hold that Level 3 AMAs are merely artefacts, tools or instruments that have no moral responsibility, and all the responsibility for what they do belongs to their developers or owners (Miller et al. 2017; Sharkey 2017). An in-between view is that AIs are not mere tools or instruments, since they make some decisions on their own, which means their developers and owners are not fully responsible for what they do, which leaves a moral "responsibility gap" in which no one is fully responsible for how such AMAs act (Gunkel 2017, p. 5).

There are three sorts of concerns this literature raises: misdirecting blame toward robots, having no one to blame, and being unable to blame robots. The first concern is based in emerging evidence that we tend to treat robots *as if* they are morally responsible and have full agency even when they clearly do not (Voiklis et al. 2016). This can lead us to misdirect our blame toward robots when blame should be directed toward the developers or owners of the robot (Bryson 2018). Second, due to responsibility gaps, when things go wrong there is no one to hold *fully* responsible for the outcomes. This concern has been raised in the context of the use of killer robots, where it has been argued that using "a weapon without a clear chain of accountability is not a moral option" (Sharkey 2012, p. 791). Third, it is unclear if it is possible to hold robots morally responsible because the possibility of punishment is central to holding others responsible, and it is unclear how we could punish robots (Wallach and Allen 2009).

In response, we do not need to blame the wrong target. There are three cases here. An AMA could be (1) fully, (2) not at all, or (3) partially morally responsible for its actions. In the first case, we should treat AMAs accordingly. While human punishments might not make sense for such bots, we could still (if it is appropriate) have reactive attitudes toward them, such as blame, and refuse to cooperate with them or give them what they want. In the second case, we should not treat them as if they *really* are morally responsible, although it might be pedagogically useful for us to *pretend* that they are responsible to promote their improvement (Sharkey 2017). But even when we are pretending in this way, we should not forget that their developer, owner, or user (depending on the case) is *really* morally responsible and should be held to account. In the third case, there are genuine responsibility gaps to bridge. These gaps can be bridged by regulation or convention. For example, if Level 3a AVs create a responsibility gap, then that gap can be filled by regulation that makes the manufacturer or owners *take* full legal responsibility for their AVs. Such a scheme has been proposed by Hevelke and Nida-Rümelin (2015) who argue that it is fair to impose a mandatory tax or insurance policy on all users of AVs to cover the risks imposed on others by such vehicles. A worry with such collective approaches to bridging the responsibility gap is that they may leave a "retribution gap" (Danaher 2016, p. 299) which arises because we want to see particular *individuals* punished when someone is harmed by a bot, such as an AV, and a mandatory *collective* tax or insurance policy on users of such vehicles will not satisfy some people's desire for retribution (Nyholm 2018). However, AMAs force us to move beyond simplistic models of responsibility that focus on single isolated actors and instead require us to embrace a complex model of diffuse "responsibility networks" (Nyholm 2018). Here we can think of the analogous case of a dog owner who can be held legally and even morally responsible for the actions of their dog, even though they do not have full control over their dog (Nyholm 2018). AMAs might also be a case where we can assign partial individual responsibility to owners or developers, as part of broader responsibility networks, even though they do not have direct control in real-time over their autonomous bots' actions. While these responses do not fully resolve the complex debate around the responsibility of AMAs, they do help us to see responsibility as a practical problem that we can solve through regulation or convention, rather than an insurmountable impediment to the creation of AMAs.

4 The case for AMAs

We have now responded to the negative case against developing AMAs, but we still need to make the positive case. To do that, we shall consider each of the seven reasons in favour of AMAs that Wynsberghe and Robbins (2019) structure their paper around rebutting. We argue here that their rebuttals are unsuccessful.

4.1 Inevitability

The first reason that Wynsberghe and Robbins (2019, p. 720) explore in favour of AMAs is that they are "inevitable" as market forces will demand the increasing use of bots in "morally salient contexts", such as healthcare, childcare, and the military, and bots will need to become AMAs to operate in such contexts. While the inevitability of an outcome is not itself a reason in favour of it, we can rephrase this as the claim that we should not oppose AMAs since they are inevitable. In response, Wynsberghe and Robbins (2019) set up a challenge. They agree that it is inevitable that bots will increasingly be placed in morally salient contexts. However, they reject that it is inevitable that bots will become AMAs just because they are placed in such contexts. They argue that almost *every* machine could potentially harm humans, including a dumb toaster, and, therefore, almost *every* machine operates in a morally salient context. But if operating in a morally salient context makes it inevitable that a machine becomes an AMA, and almost every machine operates in such contexts, then we get the supposedly "untenable" conclusion that almost every machine "must be developed as an AMA" (Wynsberghe and Robbins 2019, p. 724).

In response, we agree that the fact that a dumb toaster could burn you is not enough of a reason to make it into an AMA. We just need the toaster to work reliably and safely (i.e., remain a Level 2 AMA), not to make and act on autonomous moral judgments (i.e., become a Level 3 AMA). While we, therefore, join Wynsberghe and Robbins in rejecting the view that *all* machines that could harm humans should become AMAs, we reject their further claim that *no* machines that could harm (or fail to help) humans should become AMAs. Instead, we argue that *some* machines that could harm humans through action or inaction *should* become AMAs, and that *other* machines, such as a dumb toaster, should *not*. The task for such a nuanced position is to give an account of *when* and *in what contexts* the use of AMAs is morally appropriate. For example, Wynsberghe and Robbins (2019, p. 724) give the example of Corti, an AI which analyses breathing and speech patterns to advise emergency phone operators about the likelihood that a caller is at risk of a heart attack. But a human remains in the loop at all times, and thus Corti is not a Level 3a AMA as Corti merely advises humans and does not make autonomous moral decisions. Why not always keep humans in the loop as final decision makers? This is an important question, but the answers we give to it need to depend on the use and context. One example that we have focused on throughout the paper is that of AVs. The reason that a human cannot be kept in the loop in such cases is because it is impossible to do so in emergency braking situations, where important moral decisions must be made autonomously by the AV as there is insufficient time to bring a human into the loop (even if we wanted to). We consider some further examples like this in the next section. In contrast, there are other cases where we may want to keep a human in the loop, such as autonomous weapons systems designed to kill people (Roff and Danks 2018). The complex task, which our account lays the conceptual groundwork for, is to offer a systematic account of when humans should and should not be left out of the loop when machines make moral judgments. This allows us to take a more nuanced position than both the blanket rejection of all AMAs that Wynsberghe and Robbins defend and the blanket acceptance of all AMAs that they fear.

4.2 Preventing harm

The second claim in favour of AMAs is that we need to develop them to prevent robots from harming us (Wallach and Allen 2009). In response, Wynsberghe and Robbins (2019, p. 722) argue that making *safe* machines, not *moral* machines, is the way to prevent that harm from occurring. For example, an automatic elevator door could cause harm by closing on us. But to prevent that harm we should make it safe by incorporating a sensor so that it does not close on people, rather than transform it into an AMA.

However, Wynsberghe and Robbins have set up a false dichotomy between safety and moral agency. While safety should be a key concern, the problem, as we noted in Sect. 3.3, is that sometimes safety is not enough. This is clear in trolley-style cases in which an AV must weigh the safety of one person against the safety of another. An appeal to morality and not safety is needed to resolve such a conflict. Another reason why a focus on safety alone is too limiting is that it leads us to focus only on ways for bots to avoid causing harm through *action*, such as an elevator door closing on someone, but ignores cases of bots allowing harm to occur through *inaction*. While a safe bot might not, for example, accidently push you into the river, unlike an AMA it will not try to save you if you do fall in. Finally, a focus on safety is also insufficient when safety conflicts with other important moral values. For example, a bot designed only with safety in mind might try to ensure that a patient always takes their medication since this is the safest option, but this ignores the patient's autonomy in cases where they choose not to take their medication. To resolve such conflicts in a morally appropriate manner, we need an AMA, such as EthEl (Anderson and Anderson 2007), that can weigh up the potential for harm with respect for the patient's autonomy, and not a bot that is designed only to be safe. Generalising, we may need AMAs and not merely safe bots in cases: (1), where inaction will allow harm (e.g., failure to rescue cases), (2) when the safety of two or more parties must be weighed up (e.g., in trolley or robotic triage nurse cases), or (3) when safety is in conflict with other important values such as autonomy (e.g., autonomous refusals to take medicine cases); *and* off-loading moral judgments to humans is impossible, too inefficient, too slow, or for some other reason unnecessary or inappropriate. A dumb toaster, since it is very unlikely to be able to rescue people or face trolley dilemmas, does not need to become an AMA. A safe toaster is enough. In contrast, AVs will face trolley dilemmas and carebots and robotic triage nurses will need to weigh up safety and other important values in real-time and for this reason these bots, unlike our dumb toaster, should (eventually) become AMAs. While not *all* machines that could harm us should become AMAs, *some* should.

4.3 Complexity

Robots are becoming increasingly complex, which means it will no longer be possible to know what they will do in novel situations. Due to this complexity, we need to equip robots with "moral competence in order to govern" their "unpredictable actions in the inevitably unpredictable and unstructured human environments" in which they will operate (Wynsberghe and Robbins 2019, p. 726). Wynsberghe and Robbins respond to this challenge by claiming that we should restrict the context in which bots can operate, rather than equip them with moral competence. They give the example of AlphaGo. AlphaGo is very complex and its designers cannot predict the next move it will make in a game of Go. However, this is not a "moral problem because the context (the game of GO) is restricted" (Wynsberghe and Robbins 2019, p. 726).

While we can restrict AlphaGo to playing Go to avoid any ethical issues, restricting contexts will not always work as a solution. An obvious counterexample is that of AVs. AVs will face novel and unpredictable situations. Some of these will be ethically significant as AVs are designed to work in contexts where they have the potential to harm people very seriously. The only way to restrict AVs to a context where, like AlphaGo, they cannot harm us, is not to build or use them. But there are strong moral arguments for using AVs, including their potential to save many lives (Lin 2015). Similar considerations apply to carebots, companion bots, military robots, and general-purpose robots. We cannot put barricades around such bots or restrict them to playing online boardgames. However, in response to this point Robbins (2020, p. 394) introduces the idea of the "envelopment" of AI. For example, a dishwasher "envelops" the washing of dishes within a box, which avoids many of the issues that a humanoid dishwashing robot would face, such as accidently hitting a person while washing dishes. For Robbins (2020), we envelop an AI when we know its training data, inputs, functions, outputs, and boundaries. The only way to envelop an AV effectively, Robbins (2020, p. 394) suggests, is to put it in the equivalent of a box that is separate from everything else, as this is the only way to prevent all unexpected inputs, such as bad weather obscuring road signs or pedestrians wandering in front of it. The "box", in the case of AVs, would involve designated enveloped zones with special road signs and road markings that could be read in any weather by the AV, and no human drivers, pedestrians or cyclists allowed in the "box", since they are too unpredictable. Envelopment would certainly make the design of AVs simpler, but requiring it might be ethically problematic, since it could make us less safe overall as it would constrain the use of AVs to tightly regulated envelopment zones and force us to be at the mercy of more dangerous human drivers everywhere else. Furthermore, even within envelopment zones, AVs could still face real-time moral decisions about whose safety to prioritise as they could still face trolley-style crash scenarios and unpredictable inputs, such as when an AV breaks down suddenly in front of another AV or a person or animal breaks through envelopment zones. Since containment or envelopment is unlikely to always be successful and requiring containment can be ethically costly as it can greatly limit the beneficial uses of some AI technologies, containment will not always avoid the need for some bots to make autonomous ethical decisions.

4.4 Public trust

As many a sci-fi movie attests, there is public fear of robots and AIs. Making them into AMAs with internal ethical constraints is one way of alleviating those fears and helping to develop public trust in this new technology. Wynsberghe and Robbins' (2019) response is to differentiate between trust and reliance. We can trust people because we feel let down if that trust is betrayed. But can we *trust* machines? Or do we merely *rely* on them? And *who* or *what* are we being asked to trust: "the algorithm directing the robot; the designer; or, the development process?" (Wynsberghe and Robbins 2019, p. 728).

Depending on the complexity of the bot in question, we can be asked to trust all three: the bot, the designer, and the development process. Is this a problem? There are clearly no conceptual difficulties in trusting the last two on that list, since they include persons who we can feel betrayed by. But can we trust bots themselves? Trust is often understood to have at least two dimensions (Roff and Danks 2018, p. 6): the trust we have in machines, artefacts and strangers, which is largely a matter of predictability and reliability (e.g., when I *trust* that my car will start tomorrow), and an interpersonal trust that depends on understanding why someone acts as they do (e.g., when I *trust* that my friend will help me because she cares about me). We can clearly trust AMAs along the first trust dimension, just as we can trust toasters or cars. However, it is unclear whether making machines into AMAs will increase or decrease this dimension of trust since, as Roff and Danks (2018) argue, the more autonomous a system is, the less predictable it can be in novel circumstances. For example, I might be able to predict, and thus highly trust, what a dumb bomb will do (i.e., it will explode if it does not malfunction), but an autonomous weapons system with inbuilt ethical constraints might be harder to predict and thus trust (i.e., it might not explode if it detects that a child is present). We can also trust AMAs along the second trust dimension, since we can understand them as acting from certain moral principles or values (Roff and Danks 2018). Furthermore, as we have seen in Sect. 3.9, people *do* blame robots and hold them responsible, so it follows that we *can* trust them and feel let down if they betray that trust. But such trust typically builds up over time through repeat interactions, and thus whether, and how much, trust develops depends on various factors. For example, are those working with AMAs given special training to understand what ethical principles they act on, and are well-designed trust-building exercises undertaken between AMAs and end-users (Roff and Danks 2018)? Furthermore, making a bot into an AMA could decrease trustworthiness in some cases, since AMAs might be developed to deceive others for moral reasons (which decreases trustworthiness), such as a carebot deceiving a dementia patient to get her to take her medicine (Arkin et al. 2012). While trust of AMAs across both dimensions is thus possible, whether making machines into AMAs increases or decreases trust depends on which dimension of trust we focus on and the ways that AMAs are deployed, developed, and used. Thus, while considerations of trust do not speak for or against AMAs per se, they do force us once again to think carefully about the contexts in which we design, deploy, and use them.

4.5 Preventing immoral use

Technology has always been used to do bad things and so we should make machines into AMAs to prevent this from happening. Wynsberghe and Robbins (2019, p. 729) respond to this argument by questioning whether we know what counts as good or bad uses of technology. They use several examples to make their point. For example, consider a car with a breathalyser that will not allow its engine to start if the driver has too much alcohol in their system. But what if a woman is trying to flee domestic violence after having a few drinks? Should the car start then? Or consider the example of an elderly woman at home who wants her robot to fetch her another alcoholic drink. Should the robot do this, even though it could lead to bad health outcomes for the woman?

The worry these examples raise is that sometimes it is unclear what counts as "immoral" uses of technology, and since we do not know, then we either cannot or should not develop AMAs. Consider the first example. A dumb car cannot even consider any of the ethical issues at stake. It simply will not work if it detects enough alcohol in the woman's blood (if that is how it has been pre-programmed). Is that really better than an AMA that can adapt and try to judge what is ethical to do in that case? In the second example, consider a human relative in this scenario. Should the relative get the elderly woman another drink? It depends on several factors, such as how intoxicated the woman is. But usually the answer is "yes"—we do not normally override the choices of others to enforce optimal health outcomes on them as this would disrespect their autonomy. Anderson and Anderson's (2007) EthEl bot deals with a similar moral terrain with patients who exercise their autonomy by refusing to take medication even though this might be harmful to them. Faced with such a scenario an AMA, such as EthEl, is in a similar, if not better position, than many humans faced with the same scenario, since many humans are uninformed about what the relevant ethical principles are and how to weigh them up, whereas a bot such as EthEl is coded to take them into account and is guided by expert judgment. In contrast, a dumb bot will be programmed to operate in a certain basic way without being able to adapt or to consider explicitly *any* of the relevant moral considerations. How is the latter a better moral outcome? Of course, these examples raise questions about *what* ethical settings AMAs should have, such as whether to value autonomy over well-being, and *who* gets to decide what those settings are. But these issues are complexities to be dealt with rather than reasons not to develop AMAs per se.

4.6 Machines are better than us

Another reason given in favour of AMAs is that they will become *better* than us at morality. They might be better than us because they can process information more quickly, "be impartial, unemotional, consistent, and rational every time" (Wynsberghe and Robbins 2019, p. 729), and never get tired. Dietrich (2001) argues that robots will also be free of the morally worst evolutionary traits of humanity, such as racism and sexism. Arkin makes similar claims about the benefits of military bots since, unlike human soldiers, they will not "rape or pillage" (Wynsberghe and Robbins 2019, p. 729). Wynsberghe and Robbins (2019, p. 730) develop three distinct responses to this point. First, they argue that claiming that AMAs would be morally better than us assumes moral realism, and moral realism is, they suggest, very controversial. Second, the fact that AMAs would be "unemotional" might be a negative, since recent evidence suggests that emotions play an important role in human decision making. Third, making AMAs which are better than us at morality could lead us to outsource morality to machines and this would cause our moral skills to atrophy.

On the first concern, as we have seen above, we sometimes need AMAs because machines will be placed in situations where a moral decision *must* be made. And in such situations, an AMA might perform better than a human. For example, compare a human and an AV who are suddenly faced with the split-second choice of swerving and running into a tree, which would kill the person in the car, or continuing forward and killing a child. Faced with such a situation, a human would act instinctually (Gogoll and Müller 2017, p. 686). A proper consideration of the various moral issues at stake is impossible for a human as there is simply no time. In contrast, an AV could have time to decide what to do while explicitly taking moral considerations into account. Since it is morally better, if possible, to explicitly take moral considerations into account before making life and death decisions, and in this scenario only an AMA could do that, it follows that the AMA could act morally better than humans in this scenario. No assumption of moral realism is needed to make this claim.

On the second concern, while emotions might be essential for good *human* decision making, it is not clear that emotions are essential for good decision making per se. While emotions might be excellent heuristics or motivators for humans, they can also lead us astray, and AMAs may be able to work very differently but just as well, if not better, without emotions. Furthermore, there is ongoing work that suggests that the functional equivalents of emotions, such as guilt, can be coded into AMAs (Arkin et al. 2012), which would allow AMAs to get the functional benefits of emotions for decision-making without having to feel anything.

The third concern, that of moral deskilling, is an important one that we have already discussed in Sect. 3.7. If AMAs become much better at morality than us, then perhaps we should off-load certain moral decisions to them, just as we should give up driving if AVs turn out to be much better at it than us. It would seem morally perverse to refuse to create better moral decision makers than us just so that we can continue to make inferior moral choices ourselves. Even so, it would be a mistake to let our moral skills atrophy and disappear. But there is no reason why the presence of AMAs should rob us of all scope to exercise and develop our moral skills. Even with AMAs, we can still exercise and develop our moral powers through our relationships with other humans, through deciding what sort of life we should live, and by engaging with morally rich forms of culture from literature to videogames. We could also develop our moral skills in virtual environments, where we could explore rich moral scenarios and exercise moral skills without real-world consequences (Staines et al. 2019). There will also be many cases, especially where time is not urgent, where we will not want to outsource our moral choices to AMAs, such as who to vote for in an election. In such cases, we might benefit from having AMAs join in moral deliberations with us as advisors, without outsourcing all the moral work or the final moral decision to them.

4.7 Understanding morality better

The final reason in favour of AMAs is that the attempt to create them would improve our understanding of morality (Wallach and Allen 2009). Wynsberghe and Robbins (2019,

p. 731) argue in response that we will not learn more about morality through machine ethics but only through studying human psychology. While it might be true that we can learn much about morality through studying human psychology, it is false that we cannot also learn something about morality through trying to develop AMAs or learn something about human psychology through building computer models (e.g., Addyman and French 2012). For example, many philosophers have questioned whether morality can be reduced to an algorithm (e.g., O'Neill 1989). Attempts to build AMAs, drawing on bottom-up, top-down and hybrid approaches, will help to answer such questions and thereby improve our understanding of morality. To give another example, Anderson and Anderson (2007, 2009) claim to have discovered a *new* principle of medical ethics through their MedEthEx bot that has never been stated explicitly before. They claim that, through machine learning, they extracted this principle out of expert judgments. This shows us how attempts to build AMAs can reveal principles that might be implicit in our judgments but which we have never explicitly stated or recognised. Anderson and Anderson (2018) have recently extended this work through GenEth, a general ethical dilemma analyser that can assist in discovering ethical principles in given domains by examining expert judgments. One might respond that we *could* have made similar progress without attempting to build AMAs. Perhaps so, but that does not mean that we *would* have made that progress without such attempts or that attempts to build AMAs will not *in fact* help us to understand morality better.

5 Conclusion

In this paper we have developed a comprehensive response to the general negative case against AMAs defended by Wynsberghe and Robbins (2019). We have done so by collating and thematising the various arguments made for and against AMAs and by developing new lines of argument. We have used this as the basis for developing an all things considered judgment in favour of a nuanced positive case for developing AMAs in certain contexts. Even so, we argue that not every machine should become an AMA and not every moral decision should be outsourced to machines. We must be careful when using AMAs in sensitive contexts and ensure that we do not let our moral skills atrophy. There are also further issues to resolve around responsibility and trust. But as important as these issues and concerns are, they tell us to be careful in our pursuit of AMAs, not to abandon the project to develop them given their many potential benefits. While we should be cautious, all things considered we have strong reasons to continue to work on responsibly developing and using AMAs in certain contexts.

References

Addyman C, French R (2012) Computational modeling in cognitive science. Top Cogn Sci 4:332–341

Allen C, Smit I, Wallach W (2005) Artificial morality. Ethics Inf Technol 7(3):149–155

Anderson M, Anderson S (2007) Machine ethics. AI Mag Winter 28(4):15–26

Anderson S, Anderson M (2009) How machines can advance ethics. Philos Now 72:17–20

Anderson M, Anderson S (2018) "GenEth." Paladyn J Behav Robot 9:337–357

Arkin R, Ulam P, Duncan B (2009) An ethical governor for constraining lethal action in an autonomous system, Technical report GIT-GVU-09-02

Arkin R, Ulam P, Wagner A (2012) Moral decision making in autonomous systems. Proc IEEE 100(3):571–589

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

Bankins S, Formosa P (2019) When AI meets PC. Eur J Work Organ Psychol 26:1–15

Bekey GA (2012) Current trends in robotics. In: Lin P, Abney K, Bekey GA (eds) Robot ethics. MIT Press, Cambridge, MA, pp 17–34

Boden M (2016) AI. OUP, Oxford

Bonnemains V, Saurel C, Tessier C (2018) Embedded ethics. Ethics Inf Technol 20(1):41–58

Bostrom N (2014) Superintelligence. OUP, Oxford

Broadbent E (2017) Interactions with robots. Annu Rev Psychol 68(1):627–652

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

Bryson J (2018) Patiency is not a virtue. Ethics Inf Technol 20(1):15–26

Chalmers D (2010) The singularity. J Conscious Stud 17(9):7–65

Danaher J (2016) Robots, law and the retribution gap. Ethics Inf Technol 18(4):299–309

Darling K (2017) Who's Johnny? Anthropomorphic framing in human–robot interaction, integration, and policy. In: Lin P, Abney K, Jenkins R (eds) Robot ethics 2.0. OUP, New York, pp 173–188

Dietrich E (2001) Homo Sapiens 2.0. J Exp Theor Artif Intell 13(4):323–328

Etzioni A, Etzioni O (2016) AI assisted ethics. Ethics Inf Technol 18(2):149–156

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

Formosa P (2017) Kantian ethics, dignity and perfection. CUP, Cambridge

Gogoll J, Müller J (2017) Autonomous cars. Sci Eng Ethics 23(3):681–700

Greene J et al (2001) An FMRI investigation of emotional engagement in moral judgment. Science 293:2105–2108

Gunkel D (2014) A vindication of the rights of machines. Philos Technol 27(1):113–132

Gunkel D (2017) Mind the gap. Ethics Inf Technol. https://doi.org/10.1007/s10676-017-9428-2

Hevelke A, Nida-Rümelin J (2015) Responsibility for crashes of autonomous vehicles. Sci Eng Ethics 21(3):619–630

Himma K (2009) Artificial agency, consciousness, and the criteria for moral agency. Ethics Inf Technol 11(1):19–29

Himmelreich J (2018) Never mind the trolley. Ethical Theory Moral Pract. https://doi.org/10.1007/s10677-018-9896-4

Laukyte M (2017) Artificial agents among us. Ethics Inf Technol 19(1):1–17

Lin P (2015) Why ethics matters for autonomous cars. In: Maurer M et al (eds) Autonomes Fahren. Springer, Berlin, pp 69–85

McCauley L (2007) AI Armageddon and the three laws of robotics. Ethics Inf Technol 9(2):153–164

Miller K, Wolf M, Grodzinsky F (2017) This 'Ethical Trap' is for roboticists, not robots. Sci Eng Ethics 23(2):389–401

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

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

Nyholm S (2018) The ethics of crashes with self-driving cars. Philos Compass. https://doi.org/10.1111/phc3.12507

O'Neill O (1989) Constructions of reason. CUP, Cambridge

Peterson S (2012) Designing people to serve. In: Lin P, Abney K, Bekey GA (eds) Robot ethics. MIT Press, Cambridge, MA, pp 283–298

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

Robbins S (2020) AI and the path to envelopment. AI Soc 35(2):391–400

Roff HM, Danks D (2018) Trust but verify. J Mil Ethics 17(1):2–20

Scheutz M (2016) The need for moral competency in autonomous agent architectures. In: Müller V (ed) Fundamental issues of artificial intelligence. Springer, Cham, pp 517–527

Scheutz M (2017) The case for explicit ethical agents. AI Mag 38(4):57–64

Sharkey N (2012) The evitability of autonomous robot warfare. Int Rev Red Cross 94(886):787–799

Sharkey A (2017) Can robots be responsible moral agents? Connect Sci 29(3):210–216

Sparrow R (2012) Can machines be people? In: Lin P, Abney K, Bekey GA (eds) Robot ethics. MIT Press, Cambridge, MA, pp 301–316

Sparrow R (2016) Robots and respect. Ethics Int Aff 30(1):93–116

Staines D, Formosa P, Ryan M (2019) Morality play: a model for developing games of moral expertise. Games Cult 14(4):410–429

Tonkens R (2009) A challenge for machine ethics. Mind Mach 19(3):421–438

Tonkens R (2012) Out of character. Ethics Inf Technol 14(2):137–149

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

Turkle S (2011) Alone together. Basic Books, New York

Vallor S (2015) Moral deskilling and upskilling in a new machine age. Philos Technol 28(1):107–124

van Wynsberghe A, Robbins S (2019) Critiquing the reasons for making artificial moral agents. Sci Eng Ethics 25:719–735

Voiklis J et al (2016) Moral judgments of human vs. robot agents. In: 25th IEEE international symposium on robot and human interactive communication, 775–780

Wallach W (2010) Robot minds and human ethics. Ethics Inf Technol 12(3):243–250

Wallach W, Allen C (2009) Moral machines. OUP, Oxford