Philosophical Studies
https://doi.org/10.1007/s11098-023-01990-y

Autonomised harming

Linda Eggert[1]

Accepted: 19 May 2023

Abstract

This paper sketches elements of a theory of the ethics of *autonomised* harming: the phenomenon of delegating decisions about whether and whom to harm to artificial intelligence (AI) in self-driving cars and autonomous weapon systems. First, the paper elucidates the challenge of integrating non-human, artificial agents, which lack rights and duties, into our moral framework which relies on precisely these notions to determine the permissibility of harming. Second, the paper examines how potential differences between human agents and non-human, artificial agents might bear on the permissibility of delegating life-and death decisions to AI systems. Third, and finally, the paper explores a series of resulting complexities. These include the challenge of weighing autonomous systems' promise to reduce harm against the intrinsic value of rectificatory justice as well as the peculiar possibility that delegating harmful acts to AI might render ordinarily impermissible acts permissible. By illuminating what happens when we extend normative theory beyond its traditional boundaries, this discussion offers a starting point for assessing the moral permissibility of delegating consequential decisions to non-human, artificial agents.

Keywords Ethics of AI · Non-consequentialist ethics · Ethics of harming

1 Introduction

As we move towards a world in which autonomous systems are capable of independently inflicting harm on persons, autonomous weapon systems have been hyped up as apocalyptic 'killer robots' while questions about the ethics of autonomous vehicles have frequently been reduced to trolleyology writ large. Arguments against developing and deploying such technologies typically warn that killer robots may go rogue and that self-driving cars' algorithms will calculatedly sacrifice some people to save others. Cases in favour of developing and deploying such systems standardly appeal to their ability to save lives and reduce harm: autonomous weapon systems

✉ Linda Eggert
linda.eggert@philosophy.ox.ac.uk

[1] Faculty of Philosophy and Balliol College, University of Oxford, Oxford, UK

Published online: 12 July 2023

may decrease risks of harm to both civilians and military personnel. Self-driving cars, meanwhile, may radically reduce the number of deadly traffic accidents.

This paper explores different terrain. It asks what considerations, other than the significant potential to do good and cause harm, bear on the ethics of delegating decisions about whether and whom to harm to artificial intelligence (AI).[1] My aim is to map uncharted territory concerning (i) the extent to which familiar principles governing the permissibility of harming for human agents extend to non-human, artificial agents and (ii) how potential differences between human agents and non-human, artificial agents might bear on whether it is permissible to delegate life-and-death decisions to algorithms. Far from providing a comprehensive account of the ethics of *autonomised* harming, this paper seeks to advance the debate by illuminating unappreciated moral complexities that may weigh against familiar consequentialist considerations in favour of autonomising harmful agency. In the process, it seeks to advance debate on a number of foundational issues that must ultimately inform any sound theory of the ethics of autonomised harming. In this, the paper's mission is as much agenda-setting as it is clarificatory.

Section II discusses one distinguishing challenge for moral theory arising from the possibility of autonomised harming. Unlike human moral agents, current AI systems lack the capacity to comprehend distinctly moral features of decisions about whether and whom to harm, and the desire to respond to these features adequately. Section III examines to what extent the permissibility of autonomised harming may nonetheless be determined by the same principles as the permissibility of harms caused by human agents; and discusses the possibility that there may be unique properties pertaining to autonomised agency, which might render it a *sui generis* area of moral activity. Section IV addresses the widespread worry that delegating harmful acts to AI will create 'gaps' in accountability. It argues that the most vexing challenges arise not from the mere existence of such gaps, but from trade-offs between the moral value of harm reduction on one hand and the demands of rectificatory justice on the other. Section V discusses broader implications of the possibility of putting AI in charge of situations in which people's rights are at stake. Section VI concludes.

A few clarifications before we proceed. First, I will say 'autonomous systems' to jointly refer to autonomous vehicles and autonomous weapons. My concern is not with the ethics of *automated* harming, or with the ethics of harming 'by remote control,' but with what we might call *autonomised* harming. When harming is autonomised, the decision about *whether* to harm, and *whom* to harm, is made by an algorithm, not a human agent. No universally agreed definition of autonomy has so far emerged in debates about autonomous vehicles and autonomous weapon systems. I use the label 'autonomised' to capture two ideas: first, that autonomous systems are *made* to be autonomous; and, second, for the sake of argument, that autonomous

[1] This makes this a paper in broadly non-consequentialist ethics, to which I fully admit, though I offer no defence of the virtues of non-consequentialism here. For that, see, for example, Frances Kamm, *Intricate Ethics: Rights, Responsibilities, and Permissible Harm* (OUP, 2007).

systems' agency is nonetheless, fundamentally, *theirs*.[2] Rather than moral autonomy in the broadly Kantian sense of acting out of respect for one's moral duties, the relevant sense of autonomy is to be understood as a significant degree of independence from human agency, as a result of a prior human decision.

Second, my concern in this paper is with what is known as *narrow* AI. Narrow AI's agency is task-specific: it is designed to perform narrowly specified tasks, such as identifying hostile targets in war. This stands in contrast to what is known as artificial *general* intelligence, which might ultimately be able to perform a wide range of cognitive tasks.[3] More generally, there is no universally agreed-upon definition of what AI is. For the purposes of this paper, relevant AI systems are AI-powered machines—self-driving cars and autonomous weapon systems—with the capacity to perform harmful actions that typically 'require cognitive functions such as thinking, learning, and problem-solving when done in intelligent beings such as humans'.[4] In this context, it is tempting to succumb to broader questions about what it means to have consciousness, whether 'responsible agency' requires more than a certain type of brain, and as what kind of 'agent' a machine might qualify. These questions loom large, but they do so in the background and remain outside the scope of this discussion.

Third, moral theorists typically distinguish between first-order and second-order questions about morality. First-order questions concern what we ought to do. Second-order questions concern the nature of morality itself, including, fundamentally, whether it is *true* that a certain act is morally right or wrong, and how we can know this. The same distinction applies when we consider moral decisions delegated to AI. We can formulate first-order questions about what an autonomous system ought to do in a particular situation and second-order questions about the nature of the moral principles that apply to AI systems, whether it is true that an algorithmic decision was the right one in any given case, and what makes it so.[5] The bulk of this paper is concerned with second-order questions about moral theory, and challenges AI might raise to traditional ways of normative theorising. Hence, this is not a paper in 'applied' ethics.

[2] If autonomous systems' agency is straightforwardly reducible to human agency, the questions they raise will be more familiar.

[3] See Alan Turing, 'Computing Machinery and Intelligence,' *Mind* 59 (1950): 433–60; John R. Searle, 'Minds, Brains, and Programs,' *Behavioral and Brain Sciences* 3 (1980): 417– 24; Ray Kurzweil, *The Singularity Is Near* (Viking Press, 2005).

[4] S. Matthew Liao, 'A Short Introduction to the Ethics of Artificial Intelligence,' in S. Matthew Liao (ed), *Ethics of Artificial Intelligence* (OUP, 2020), 1–42, 3. See also Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed (Prentice Hall, 2010), 2.

[5] This is different from what has been described as 'computational metaethics' or 'automated reasoning about automated moral reasoning.' See Gert-Jan C. Lokhorst, 'Computational Meta-Ethics: Towards the Meta-Ethical Robot,' *Minds and Machines* 21 (2011): 261–274; Bart Wernaart, 'Developing a roadmap for the moral programming of smart technology,' *Technology in Society* 64 (2021).

Finally, in charting unexplored territory, this paper offers a rough and preliminary sketch of a sprawling and dense landscape. Its aim is to articulate and illuminate issues, rather than to resolve them, and to offer a roadmap for further deliberation.

2 Agency without duties?

Existing AI systems lack the capacity to assess whether causing a certain harm is right or wrong.[6] If AI systems are to be deployed in situations that involve decisions about whether and whom to harm, one question is what principles algorithms should be programmed to follow.[7] Standard theories of defensive harming and liberal analytical theories of justice focus on human agents who possess certain capacities of moral reflection and self-control. Such theories also orient themselves in relation to individual rights. Whether it is permissible to harm and what we owe to others is typically determined with reference to the importance of *protecting* individual rights, the conditions for *forfeiting* one's rights, and the consequences of *violating* the rights of others. In most cases, the rights in question correspond with duties. Your permission to defend yourself against an attacker presupposes both your right not to be harmed and your attacker's correlating enforceable duty not to harm you.

This poses a considerable challenge for theories of *autonomised* harming. Non-human, artificial agents, which possess neither rights nor duties in any conventional sense, must somehow be integrated into a human-centred moral framework that relies on precisely these concepts to determine whether harming is permissible.

While some philosophers have explored whether we should think of robots as rights-holders, a no less important question is whether we can view them as *duty*-bearers.[8] On mainstream views of moral agency, the possession of autonomy is intertwined with being constrained by duties. Those unable to recognise others' rights and act in accordance with correlating duties, such as non-human animals, young children, or psychopaths, are not usually regarded as fully autonomous agents. Responsible agents who do violate others' rights make themselves liable to punishment, on the condition that they acted, to a certain degree, autonomously.

Now, while autonomous moral agency for human persons generally comes with duties that correspond with others' rights, autonomous systems are peculiar in that they seem to possess something like the former despite lacking the latter. Autonomised agency might thus constitute a *sui generis* area of moral activity: autonomous

[6] See S. Matthew Liao, 'A Short Introduction to the Ethics of Artificial Intelligence,' in Liao (ed), *Ethics of Artificial Intelligence*, 9.

[7] For example, Wendell Wallach and Colin Allen (eds), *Moral Machines: Teaching Robots Right from Wrong* (OUP, 2008); Michael Anderson and Susan Leigh Anderson (eds), *Machine Ethics* (CUP, 2011).

[8] Eric Schwitzgebel and Mara Garza, 'A Defense of the Rights of Artificial Intelligences,' *Midwest Studies in Philosophy* 39 (2015): 98–119; Sushma Raman and William F. Schulz, *The Coming Good Society* (HUP, 2020); John Basl and Joseph Bowen, 'AI as a Moral Right-Holder' in Markus D. Dubber, Frank Pasquale, and Sunit Das (eds), *The Oxford Handbook of Ethics of AI* (OUP, 2020); Liao, 'The Moral Status and Rights of Artificial Intelligence' in Liao (ed), *Ethics of Artificial Intelligence*. For general discussion, see Liao, 'The Basis of Human Moral Status,' *Journal of Moral Philosophy* 7 (2010): 159– 79.

systems have the capacity to act autonomously without, in doing so, being constrained by moral duties in any conventional sense.[9] Call this *Agency Without Duties*.

Agency Without Duties suggests that autonomous systems' 'autonomy' differs from that of paradigmatic human moral agents in that it is severed from the conventional constraints of moral agency. Without sentience or consciousness, and unbound by moral duties, non-human, artificial agents lack key characteristics of paradigmatic responsible agents who are subject to duties that correspond with others' rights.

Theories of liability to punishment provide a useful starting point for determining what it takes to qualify as a 'responsible moral agent.'[10] Such theories usually presume a number of psychological criteria for responsible agency. In addition to having had some particular intention or knowledge, the standard presumption is that the agent must have 'possessed certain powers of understanding and control.'[11]

Algorithms—just like non-human animals, young children, and other agents lacking full normative agency—lack 'certain powers of understanding and control' that are necessary components of responsible agency. Suppose also, as is standardly assumed, that rights not to be harmed correspond with duties not to harm, such that a right is violated when a duty is breached.

One problem arising from *Agency Without Duties* is this: if the violation of a right presupposes the breach of a duty, and autonomous systems cannot be said to possess duties, how can we say that people who, despite possessing rights not to be harmed, are harmed or killed by a self-driving car or an autonomous weapon system are victims of a rights violation?

An immediate response is that those human agents who designed an algorithm or activated an autonomous system are the relevant responsible agents, and that it is they who transgress the rights of those who are consequently harmed. But if there are cases in which matters are less straightforward—perhaps because certain autonomous systems indeed act with such independence from human agents that harms they inflict on human persons cannot be traced back to a human violation of a human duty—we need a response to *Agency Without Duties*. I will lay out three possibilities.

2.1 The bad luck approach

One possible response is that instances in which innocent persons suffer autonomised harm are morally comparable to suffering misfortune. On this view, being harmed by a self-driving car or an autonomous weapon system is similar to suffering

[9] I take it that, if autonomised agency were constrained by duties of human agents, such as designers or programmers, AI systems' actions would not be autonomous in the sense relevant to this discussion.

[10] Much has been said about the idea of 'collective' or 'corporate' agency. But this, I think, can only tell us little about autonomised agency. For a more confident view, according to which autonomous systems are, for all intents and purposes, sufficiently like corporate agents, see Michael Robillard, 'No Such Thing As Killer Robots,' *Journal of Applied Philosophy* 34 (2018): 705–717.

[11] See, for example, H. L. A. Hart, 'Postscript: Responsibility and Retribution,' in *Punishment and Responsibility: Essays in the Philosophy of Law* (OUP, 1968), 210–236.

harm through bad luck, in that it involves no wrongful agency. Call this the *Bad Luck Approach*. On the Bad Luck Approach, being struck by an autonomous vehicle or being harmed by an autonomous weapon is no morally different from being struck by lightning, getting caught in an unexpected landslide, or being attacked by a shark—all instances of being harmed without any wrongdoing.

But the Bad Luck Approach seems inadequate. Presumably, in many cases, people harmed by autonomous systems do have their rights transgressed. The only exceptions to this are cases in which people have made themselves liable to be harmed, for example by posing threats of unjust harm in war. In the absence of liability, presumably we want to be able to say that victims of autonomised harm either have their rights unjustifiably violated or, if this is justified on lesser-evil grounds, justifiably infringed.[12] The difficulty is that the peculiarities of *Agency Without Duties* call this possibility into question. Consider an innocent person who is run over by a trolley that was set in motion by the wind. Presumably, she is a victim of bad luck rather than having been wronged. In the case of autonomised harming, neither label seems accurate. Being harmed by, say, an autonomous vehicle is neither like being run over by a reckless or negligent human driver nor like being run over by a runaway trolley set in motion by the wind.

The Bad Luck Approach is thus unsatisfying. The possibility of autonomising harmful agency should not stop individuals' rights not to be harmed from serving as the main determinant of whether harming is permissible, and of whether victims of harm have been wronged.[13] Harming people who have a right not to be harmed should have weighty moral consequences, such as the incurrence of rectificatory obligations. Rights violations typically generate claims to compensation and, in some cases, liability to punishment. Harming people who have rights not to be harmed is not something a theory of the ethics of autonomised harming should be nonchalant about.

2.2 The Thomsonian approach

A second response to *Agency Without Duties* is to discard the standard assumption that rights violations presuppose breaches of duties. Since this is reminiscent of a view Judith Thomson defended a while ago, I will call this the *Thomsonian Approach*.[14] In *The Realm of Rights*, Thomson presents a rights-based account of defensive harming that assumes that rights can be violated in the absence of moral agency. This view has been widely criticised for its failure to presuppose any degree

[12] I follow the basic Thomsonian distinction between rights infringements and rights violations here.

[13] These are two distinct questions, since it is possible that the same harmful act is both permissible and wrongs the victim; it is also possible that an act of harm would not wrong the victim (for example, because she has made herself liable) but is nonetheless impermissible (for example, because it would be gratuitous). This latter case arises for views on which the necessity requirement for defensive harming is considered 'external' to liability. To avoid excessive complication, I will set such nuances aside. This is just to explain why I distinguish the question of whether harming is permissible from the question of whether a victim would be wronged by being harmed.

[14] Judith Jarvis Thomson, *The Realm of Rights* (HUP, 1992).

of moral agency on the part of the rights-violator. As Jeff McMahan put it, 'neither a falling boulder nor a charging tiger can be subject to a moral constraint; thus neither can violate a right.'[15]

But precisely this feature of Thomson's view renders it relevant to the context of autonomised harming. If we want to resist comparing being harmed by an autonomous system to, say, being hit by a falling boulder, Thomson's account may look oddly appealing: it might offer a basis on which to claim that rights can be violated through autonomised agency. It would also have the distinct advantage of providing grounds for rectificatory obligations. Victims of rights transgressions have stronger claims to compensation than victims of bad brute luck, at least so long as other things are equal. If responsible agency is not a necessary condition for the violation of a right, the Thomsonian Approach might help us make sense of how to respond to harms inflicted by autonomous systems.

But, in the end, I suspect that it will remain difficult to see how we could reconcile the assumption that rights can be violated in the absence of moral agency with the assumption that only *responsible* agents can be subject to moral constraints.[16] For the Thomsonian Approach to become applicable, we would likely have to introduce a new category of moral agency for autonomous systems that (a) recognises the ability to violate duties without (b) presupposing the ability to meaningfully discharge them. We will return to this shortly. First, consider a final response.

2.3 The moral room approach

A third possibility is that AI systems could be designed to act *as though* they were abiding by moral duties that correspond with people's rights. Call this the *Moral Room Approach,* based on John Searle's 'Chinese room' thought experiment and Mahi Hardalupas's adaptation to a 'moral room.'[17] In this version, the agent finds herself inside the room and must output 'moral decisions' addressing moral challenges, using a 'moral rulebook.' Proponents of the Moral Room Approach would say, for instance, that autonomous weapon systems could be programmed to act *as though* they 'knew' that innocent civilians have rights not to be harmed. The Moral Room Approach might even appeal to those who would prefer broadly deontological AI systems, insofar as AI systems might be programmed to act *as though* they were following, say, the categorical imperative.[18]

But the Moral Room Approach seems unsatisfying. Complying with moral duties essentially means deliberately doing what morality requires while being aware of the alternative courses of action which one eschews. One must *know* that one is morally

[15] See also David Rodin, *War and Self-Defense* (OUP, 2002), 81–83; Jeff McMahan, 'Self-Defense and the Problem of the Innocent Attacker,' *Ethics* 104 (1994): 252–290, 276; Michael Otsuka, 'Killing the Innocent in Self-Defence,' *Philosophy & Public Affairs* 23 (1994): 74–94, 80.

[16] See Thomson, *The Realm of Rights*, 77.

[17] See Searle, 'Minds, Brains, and Programs'; Hardalupas, 'A Systematic Account of Machine Moral Agency.'

[18] Thomas Powers, 'Prospects for a Kantian Machine,' *Intelligent Systems*, IEEE 21 (2006): 46– 51. Helen Frowe, *The Ethics of War and Peace*, 2nd edition (Routledge, 2015), 19.

constrained in one's actions in some way. Acting out of respect for one's moral duties, arguably, is part of what it means to be an autonomous moral agent to begin with.[19] 'Being a moral agent,' as Daniel Butt puts it in a different context, 'means being committed to the idea that justice should prevail over injustice.'[20] Even if AI systems were programmed to act *as though* they were under certain moral constraints, they would not act or make decisions 'for the right reasons.'[21] Besides, the notion of moral agency presupposes the ability to recognise particular reasons as *moral* reasons.[22]

This brings to the fore an aspect of responsible moral agency that is rarely made explicit. Some things *should* be difficult. Certain features of the moral universe, such as value conflicts and conflicts between rights, render certain acts and decisions distinctly challenging. Receptiveness to these features—that is, the capacity to recognise, comprehend, and aptly react to them—is an essential element of autonomous moral agency. Call this *moral receptiveness.* If moral receptiveness matters, the Moral Room Approach is unsatisfying.

Moral receptiveness has at least three components. The first manifests itself as the capacity to perceive moral features as such. The second is related: this is the capacity for moral reasoning, to engage in normative reflection about how different moral features of situations interact. This includes the capacity to identify values, and potential conflicts between them, the capacity to weigh different considerations against one another, and to make and justify trade-offs between them. It also includes adaptability to new situations—for example, when driving, or in the fog of war. When encountering new situations, including new moral challenges, moral agents are able to reason through them, to engage in normative reflection to determine what the right course of action is, without having 'learned' what to do in that particular situation—without, as it were, having been trained on certain 'inputs.'[23] In this sense, moral receptiveness is also related to moral understanding—the capacity to understand what *makes* certain acts right.[24]

When we use thought experiments to test our intuitions about a particular problem, we rely on precisely this capacity for moral reflection to reason our way through potentially new, unfamiliar challenges. For example, we might see a difference between (a) turning the trolley away from the five and towards the one in the standard trolley case and (b) pushing a person standing on a footbridge onto the tracks to stop the trolley from running over the five. One way of reasoning our way through

[19] Kantians may wish to mentally insert a Kant reference here, acknowledging that autonomy means acting with respect for the moral law.

[20] Daniel Butt, *Rectifying International Injustice* (OUP, 2009), 128.

[21] Duncan Purves, Ryan Jenkins, Bradley J. Strawser, 'Autonomous Machines, Moral Judgment, and Acting for the Right Reasons,' *Ethical Theory and Moral Practice* 18 (2015): 851–872.

[22] For example, Christine M. Korsgaard, *The Sources of Normativity* (CUP, 1996); T. M. Scanlon, *What We Owe to Each Other* (HUP, 1998); see also Scanlon, 'Forms and Conditions of Responsibility, in R. Clarke, M. McKenna, and A.M. Smith (eds), *The Nature of Moral Responsibility: New Essay*s (OUP, 2015).

[23] Thanks to Rob Reich for helpful discussion on this.

[24] On moral understanding, see Alison Hills, 'Moral Testimony,' *Philosophy Compass* 8 (2013): 552–559; Hills, 'The Intellectuals and the Virtues,' *Ethics* 126 (2015): 7–36.

this involves appealing to certain moral asymmetries, say between killing as a means and killing as an unintended side effect, or between doing and allowing harm. But it is neither obvious that such deontological distinctions could be translated into algorithmic codification top down nor that bottom-up approaches to machine learning could work out the difference between (let us stipulate) the *permissible* act of killing the one to save the five in the standard trolley case and the *impermissible* act of killing the one to save the five in the footbridge variation.

The third component of moral receptiveness goes beyond capacity. This is the *desire* to get moral decisions right, and scruples at the possibility that we might get such decisions wrong. This is why moral choices in the face of risk and uncertainty are particularly difficult. Moral agents have the capacity to care about doing the right thing for morality's sake. Irrespective of how we think rightness is determined, doing the right thing *matters* to moral agents. Even without a sophisticated theory of the phenomenology of moral agency, we can say that something would be amiss if one did not recognise as troubling acts that transgress fundamental rights or values. It is, quite plainly, apt for us to become frustrated by trolley cases, ticking-bomb scenarios, and Sophie's-choice type dilemmas.

While our frustration in these contexts is typically taken for granted, it seems to me that it is actually the manifestation of a vital element of responsible agency worth defending. When different values and duties pull in different directions, the sacrifice in not being able to do everything we ought to do—save all six in the trolley case, find the ticking bomb without torturing someone, Sophie saving both her children—is, in a non-trivial sense, real. Hence the significance of being receptive to the distinctly moral features of a situation.[25] If the recognition that doing the right thing matters is itself of intrinsic value, the Moral Room Approach is unsatisfying.[26]

What we should say about autonomous systems designed to act *as though* they are under moral constraints thus depends in part on the extent to which we care about moral receptiveness and moral understanding.[27] Whether a lack of moral receptiveness and moral understanding constitutes a reason against autonomising harmful acts may depend on the specific circumstances of the case. In some cases,

[25] There is no reason to think that human agents always comprehend the moral features of the circumstances in which they act. But this does not mean that it would not be better if they did. Human agents also reasonably disagree about how to assess the moral features of a situation. The point is that it matters that moral agents comprehend that something of moral interest is at stake. More on this in Section V.

[26] One question which I will not address here, but which theories of autonomised harming may need to accommodate, is that of whether there is a possibility that AI could somehow develop moral receptiveness to a similar or greater degree than human persons. Whether this could ever be in the realm of technological possibility is of secondary interest here. What matters for moral theory is whether we have grounds for thinking that only human persons, with certain human experiences, emotional and cognitive capacities, can adequately appreciate why morality and justice matter unlike anything else. So, a theory of autonomised agency might need to say something about the question of whether AI could somehow be 'better' at morality than humans, and how we, with our inferior moral minds, could then assess the quality of AI's moral judgements. Both are questions this paper sets aside, but it would be amiss not to mention, with regard to the former question, that at least some primates seem clearly sensitive to the appeal of relational egalitarianism (Maria Konnikova, 'How We Learn Fairness,' *The New Yorker* (7 January 2016), available at: https://www.newyorker.com/science/maria-konnikova/how-we-learn-fairness).

[27] Thanks to an anonymous reviewer for pressing me on this.

for example, for sentencing decisions in criminal justice, people's interest in having life-affecting decisions be made by fellow human agents with a certain capacity for moral receptiveness and understanding might be dispositive. In other cases, for example, when it comes to accurately sorting recycling, the human capacity for moral receptiveness and understanding is not a reason to limit those decisions to humans. What the circumstances are in which moral receptiveness and understanding matter is a question I will leave open. Cases involving judgements about what is right or just, and those involving some kind of moral sacrifice—such as cases in which innocent people's rights are, albeit permissibly, infringed on lesser-evil grounds—are likely among the most important ones.

So much for the first challenge, arising from *Agency Without Duties.* The three possible responses I have sketched—the Bad Luck Approach, the Thomsonian Approach, and the Moral Room Approach—are far from comprehensive. To the extent that questions arising from *Agency Without Duties* remain unresolved, the approaches' limits I highlighted point the way to further avenues of enquiry. Time now to consider the second challenge.

3 Narrow alignment and divergence

Do the same principles that govern the permissibility of harming for human agents extend to AI systems? One issue that has received considerable attention in the AI ethics literature is that of 'value alignment.' This usually comes up in the context of 'superintelligent' AI which, some people fear, might end up harming or destroying humanity unless AI's values are 'aligned' with ours.[28] In the context of narrow AI, a correspondingly narrower question is that of whether the principles AI systems are programmed to follow ought to be 'narrowly aligned' with those which we take to govern human agency in the same context. For example, should autonomous weapon systems (AWS) be made to follow the same rules that human combatants are required to follow in war? That is, do the principles governing what constitutes 'just conduct' for human combatants also determine what constitutes 'just conduct' for AWS? Similarly, is what autonomous vehicles ought to do determined by how human drivers should behave?

Morality is a comprehensive system. Its principles apply in all areas of human activity. One contested upshot of this is that the same principles apply in different contexts. For example, the same principles that govern the permissibility of defensive harming in 'ordinary' life also govern the permissibility of harming in war.[29] If the same principles extend to all areas of human activity, whether and why it is

[28] For example, Roman V. Yampolskiy (ed), *Artificial Intelligence Safety and Security* (Chapman & Hall, 2018); Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking, 2019). For a sceptical view of the possibility of superintelligence, see Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (MIT Press, 1992).

[29] This is a view not universally held. Traditional just war theorists deny that the same principles govern the permissibility of harming within war and outside war. See, for example, Michael Walzer, *Just and Unjust Wars* (Basic Books, 1977). I take it that the comprehensiveness of morality is compatible with the criticism that reductive individualism may be too reductivist and too individualist in some respects. Those who disagree, and hold that war is a special area of human activity in which 'ordinary' moral prin-

permissible to use defensive force to protect one's right not to be harmed is independent of the context in which one happens to find oneself—whether one is finding oneself at home, driving down the street in one's car, or fighting enemy combatants in armed conflict.[30]

Considering autonomous systems, one question is whether morality behaves similarly for non-human, artificial agents, such that there is no difference between the permissibility of harming in everyday traffic on one hand and the permissibility of harming in armed conflict on the other. If that is the case, the same principles that govern the permissibility of harming for autonomous *vehicles* govern the permissibility of harming for autonomous *weapons*. But that does not seem quite right.

3.1 Problems of partiality and permissiveness

We ordinarily assume that people are under no duty to promote the greater good, or the welfare of others, when this would come at significant cost to them. In the absence of special duties and liabilities, one may permissibly prioritise one's own life and goals over the welfare of others. If all one could do to save five strangers is to sacrifice one's life, generally speaking, one may permissibly choose not to do so. Though we may incur demanding duties to rescue, we do not typically have to make ourselves available as means to bringing about the greater good.

The advent of autonomous vehicles (AVs) raises the question of whether governments are permitted, or even required, to ensure that AVs minimise harms impartially, although individuals are not required to do so. Since there is no obligation on individuals to maximise overall utility, should regulators allow individuals to programme AVs not to maximise utility; or should the state prioritise its own presumed duty to make utility-maximising decisions about AI systems? This generates what I will call the *Problem of Partiality.*

One question at the heart of this problem is whether AI systems should be made to act completely impartially, as judged from the 'point of view of the universe,' to borrow Sidgwick's phrase, or whether their designers should programme partiality for AI systems' users' interests into those systems' controlling algorithmic structure. This, in turn, raises the question of who has the legitimate authority to programme and set vehicles' risk-distribution algorithms: manufacturers, the state and its regulators, or individual citizens? How we answer this question may affect what algorithms should be optimised for, given that individuals may, within limits, have agent-relative prerogatives to protect themselves at the expense of others, whereas states are obligated to regard all citizens' lives as being of equal value. If this is right, the Problem of Partiality points towards a discrepancy between what individuals are

Footnote 29 (continued)

ciples are abrogated, may find the problem I discuss here less troubling than those who believe that the principles of defensive harming are the same in both contexts.

[30] Bear in mind that my concern here is with moral rather than legal principles. My insisting that morality must be comprehensive is compatible with the assumption that the law must differ from morality in certain cases, including, in some regards, in armed conflict.

permitted to do and what their AVs could permissibly be programmed to do in the same circumstances.

My aim here is not to resolve the Problem of Partiality.[31] What matters, for our purposes, is this: if AVs must be prohibited from performing acts on behalf and in the interest of their passengers which it is permissible for those individuals to perform themselves, there is a clear divergence between the principles that govern individuals' actions and those that apply to AVs.[32]

Consider now autonomous weapon systems (AWS). Unlike human persons, AWS are not subject to moral powers, they cannot be thought to be blameworthy, morally responsible, or excused for harming people. By contrast, many human combatants are at least partially excused for the threats they pose, for example as a result of operating under severe epistemic constraints or duress, which makes it impossible for them to carefully assess, and act in accordance with, their own and others' moral rights and duties.[33]

Insofar as human combatants may be absolved from blame for harming morally innocent individuals, whereas nothing of the sort is true for robots, it seems that it must remain impermissible for AWS to do what it may be excusable for human combatants to do in the same circumstances. Most significantly, this applies to the use of lethal force against persons who have rights not to be harmed. If even such a cursory sketch captures something of significance, the principles governing just conduct for AWS must be considerably more restrictive than those governing just conduct for human combatants.

At least some of those considerations that permit human persons to exercise partiality in determining how to distribute risks of harm and that excuse individuals for harming others in certain circumstances do not plausibly extend to non-human, artificial agents. Consequently, *narrow alignment* would render the principles governing harmful acts by autonomous systems unduly permissive. It would therefore be a mistake to assume that autonomous systems' algorithms must be programmed to behave according to the same permissions and constraints that govern human agents in the same circumstances.

3.2 Double divergence

Insofar as AVs' crash avoidance features must aim to minimise harms overall, and AWS must identify and engage only liable targets, these autonomous systems' objectives are fundamentally different. The principles governing the permissibility of

[31] For broader discussion, see Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, 'The Social Dilemma of Autonomous Vehicles,' *Science* 352 (2016): 1573–76. On my view, risks of being harmed as a side effect of impartial collision-avoiding software may be justified through a combination of ex ante contractualist and fair-play theories—assuming, roughly, that imposing certain risks on persons is justifiable when they stand to benefit from this and would reasonably consent to bearing this risk.

[32] This may just be a manifestation of the ubiquitous tension between individual rights and aggregate welfare or the greater good. Insofar as this issue is exacerbated by AVs, I take it, my point stands.

[33] For example, Jeff McMahan, *Killing in War* (OUP, 2009).

harms caused by AVs must differ from those governing the permissibility of harms caused by AWS. Both sets of principles also diverge from those that apply to human agents in the same circumstances.[34] The principles determining what harms human agents may permissibly cause are the same no matter the context; but the principles determining what harms an AV may permissibly cause differ from those determining what harms a human driver may permissibly cause in the same situation, and the principles determining what harms an AWS may permissibly cause differ from those determining what harms a human combatant may permissibly cause in that situation. Call this *Double Divergence.*

Double Divergence is the observation that the principles governing the morality of autonomised harming differ from those governing the morality of harming for human agents in at least two ways. First, they broadly diverge regarding the *site* on which they apply: the same principles govern the permissibility of defensive harming for humans in both ordinary everyday life and in armed conflict, whereas different principles govern the permissibility of harming for AVs on one hand and AWS on the other. Second, these principles more narrowly diverge regarding their *content*, insofar as (i) the principles governing the permissibility of harms caused by AVs differ from those governing the permissibility of harms caused by human drivers in the same circumstances, and (ii) the principles governing the permissibility of harms caused by AWS differ from those governing the overall permissibility of harms caused by human combatants in armed conflict. Hence Double Divergence.

Insofar as autonomised agency is governed by principles that behave differently from those governing the ethics of harming for human persons, we face the question of whether we should treat autonomised agency as a *sui generis* area of moral activity. If there are properties pertaining to the ethics of autonomised harming that are unique to AI systems, their algorithms may need to follow principles that behave fundamentally differently from those that govern the permissibility of harming for paradigmatic human agents.

Here is one way to think about this. In addition to the traditional deontic categories of ‘doing’ and ‘allowing,’ or ‘acts’ and ‘omissions,’ we might conceive of a new mode of agency, that of ‘autonomising.’ The notion of *autonomising* might combine the moral weight we typically ascribe to instances of *doing* harm with the acknowledgement typically accompanying the notion of *allowing* harm, namely that there is a difference between what we *do* and what *happens*.

Perhaps there is nothing morally mysterious about autonomised agency, since algorithms are designed by human agents. Given that AI systems are artefacts created by human persons, one might think, no new concepts are needed, especially if human agents will be accountable for harms caused by autonomous systems. But neither the fact that human agents design the technology nor the possibility that human persons may, in some sense, be held responsible for the conduct of autonomous systems resolves the question we set out to address. What principles

[34] More precisely, we might say that the same principles apply, but that—given differences in types of agency—the permissions they yield differ for human and non-human, artificial agents. But this difference is merely semantic. The idea remains the same.

autonomous systems should follow in determining whether and whom to harm, and whether those principles are the same as those that apply to human agents, is orthogonal to whether the autonomisation of agency creates responsibility 'gaps.' The possibility of holding human agents accountable neither presupposes that we know which principles apply to autonomous systems nor entails that the same principles apply to AI systems that apply to humans. This brings us to the third challenge.

4 Accountability

Concerns about accountability are ubiquitous in debates about the ethics of autonomous systems. Debates about AI ethics invariably bring up the question of who is ultimately 'responsible' for harms caused by AVs or AWS.[35] That algorithms cannot apologise, be blamed, interrogated, or held responsible for harms wrongfully caused in any conventional way is a feature many people find troubling.

Whether, and why, the possibility of a 'gap' might worry us depends on what precisely we are after in enquiring about responsibility or accountability for autonomised harms.[36] For example, the concern might be with fairly distributing (compensatory) burdens in mitigating harms caused by AI systems. Or the aim might be to work out who, if anyone, is liable to punishment in any given case. Responsibility matters in both contexts, but for different reasons. What follows is a starting point for how we might advance debates about rectificatory justice in the context of autonomised harming.

4.1 The gap

Mainstream positions fall within three broad categories. The first insists that the autonomisation of harming necessarily results in a 'responsibility gap.'[37] As a result, neither persons nor AI systems can be held meaningfully responsible for harms inflicted by the latter, which renders their use *pro tanto* impermissible. The second position suggests that autonomised harms *can* be traced back to human agency to a sufficient degree to hold persons morally responsible for harms inflicted

[35] There are sophisticated approaches for working out questions like to what extent we are responsible for what others do as a result of what we do, or how compensatory burdens for harms ought to be distributed in the absence of individual moral wrongdoing. To name just a couple of examples, see John Gardner, *Offences and Defences: Selected Essays in the Philosophy of Criminal Law* (OUP, 2007) for the former and David Miller, 'Distributing Responsibilities,' *Journal of Political Philosophy* 9 (2001): 453–471, for the latter.

[36] One difficulty is that the related notions of 'accountability' and 'responsibility' are both highly complex and used in different ways by different people. Rather than proposing distinct definitions here, I will appeal to both notions, as potential distinctions have no bearing on my argument.

[37] Rob Sparrow, 'Killer Robots,' *Journal of Applied Philosophy* 24 (2007): 62–77; Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata,' *Ethics and Information Technology* 6 (2004): 175–83.

by autonomous systems.[38] The third position suggests that the problem of attributability for harms caused by autonomous systems is no different from any ordinary collective action problem, and that we should, accordingly, treat AI systems like social institutions.[39] What unites these three positions is their assumption that the permissibility of using autonomous systems will depend, in part, on the possibility of accountability. Call this the *Accountability Condition*: the possibility of accountability is itself a condition for the permissibility of autonomised harming. In what follows, I suggest that the three main views either can have their objections met or underestimate the degree to which autonomous systems act independently of human agents.

4.1.1 Compensation without moral responsibility

First, suppose that 'responsibility gaps' are real in the sense Rob Sparrow describes. This need not pose an insurmountable obstacle, depending on which demands of rectificatory justice we are aiming to meet. There are many potential routes to remedial obligations that do not involve the kind of moral responsibility that we ascribe only to responsible moral agents. Insofar as remedial obligations may arise from, for example, mere causal responsibility, communal membership, or having benefited from injustice, the absence of moral responsibility for harms inflicted by autonomous systems need not preclude the possibility of rectificatory obligations. We might compare these challenges to common situations in which perpetrators are no longer alive or otherwise unable to compensate their victims for wrongful harms. Victims may still possess claims to compensation even if these duties cannot be discharged by those morally responsible for these harms.[40]

While moral responsibility is a condition for culpability, which is a condition for liability to punishment, moral responsibility is not a condition for liability to compensation. Compensatory obligations for harms caused by autonomous systems may arise even if neither algorithms nor persons occupying some role in the causal architecture of the harmful action in question are *morally* responsible for the harm caused. If this is right, moral responsibility gaps might exist but, so long as we are not concerned with punishment, need not trouble us. So, to the extent that rectificatory justice in the absence of individual moral responsibility is possible, the fact that the autonomisation of harming might result in responsibility gaps may be less troubling than typically assumed.

[38] Tom Simpson and Vincent C. Müller, 'Just War and Robots' Killings,' *The Philosophical Quarterly* 66 (2016): 302–322.

[39] Michael Robillard, 'No Such Thing As Killer Robots.'

[40] For example, David Miller, 'Distributing Responsibilities'; Butt, *Rectifying International Injustice* and 'Inheriting Rights to Reparation: Compensatory Justice and the Passage of Time,' *Ethical Perspectives* 20 (2013): 245–269; Anna Stilz, 'Collective Responsibility and the State,' *Journal of Political Philosophy* 19 (2011): 190–208; Robert E. Goodin and Christian Barry, 'Benefiting from the Wrongdoing of Others,' *Journal of Applied Philosophy* 31 (2014): 363–376.

4.1.2 The Acceptance View

Another possible approach to addressing responsibility 'gaps' is that people might voluntarily *accept* responsibility for harms their AI systems might cause.[41] Call this the *Acceptance View*. For example, somewhat like signing a waiver, those using self-driving cars or AWS might commit to accepting responsibility for any harms the AI system might cause. The primary function, presumably, would be to assume remedial responsibilities, to mitigate harms that might be caused. One ground for such acceptance of responsibility is benefit, as per the 'beneficiary pays principle' from other contexts of compensatory justice. Expected benefit, we might assume, constitutes a plausible ground for assuming responsibility for autonomised harms, especially if the AI systems in question are programmed to prioritise their users' interests over other people's.

Detailed discussion of the Acceptance View would go beyond the scope of this paper. Questions it raises include whether accepting responsibility in this way could be a duty and, if so, whether it is enforceable. One likely difficulty is that the beneficiary pays principle typically applies in cases in which someone benefited unjustly, or we must distribute compensatory burdens between non-liable people, and benefit is one aspect in which things are not equal between them. So, it may not be clear whether we can appeal to the same principle in this context, especially since the expected benefit of reducing harms may be what justifies the use of AI systems in the first place.

4.1.3 The fault view and social institutions

Consider now the two other positions I outlined, which claim that responsibility *can* be traced back to human agents or that AI systems are relevantly similar to social institutions.[42] One version of the former, which we might call the *Fault View,* is as follows. Human agents might be at fault and might be held responsible for harms caused by AI systems, either as a result of recklessness or negligence.[43] If the human agent activating, say, an AWS was aware that it might cause harms that should not be caused, then the human agent might have been reckless in deploying that AWS. If the human agent, on the other hand, was *not* aware of this risk, then they might have been negligent in deploying that AWS.[44]

[41] For a view along these lines, see Maximilian Kiener, 'Can we Bridge AI's responsibility gap at Will?' *Ethical Theory and Moral Practice* 25 (2022): 575–593.

[42] As an anonymous reviewer points out, we might respond to Agency Without Duties by saying that autonomised harms are the result of a failure, on the part of individuals or institutions, to take appropriate measures to prevent those harms. A similar thought seems to me at work in common calls for 'meaningful human control,' especially in the military context. On my view, human control may render harms caused by AWS instances of wrongful omissions by human agents who failed to intervene and prevent those harms.

[43] Strict liability is a third option, which I set aside here.

[44] Based on a comment by Stuart Russell at Wadham College, Oxford, several years ago.

One challenge for both the Fault View and views that compare autonomised agency to that of social institutions is that these approaches risk underestimating the degree to which autonomous systems' agency is theirs; that, as Sparrow put it, 'their actions originate in *them* and reflect *their* ends'.[45] Although algorithms are deterministic and act in accordance with set rules, they are neither necessarily predictable nor controllable.[46] And even if autonomous systems' agency may, to a certain degree, be comparable to that of psychopaths, child soldiers, and social institutions, comparisons to collective responsibility or non-responsible threats may not adequately address the fact that autonomy in AI systems is a non-binary feature. It does not either exist or not exist.[47] Rather, autonomous systems will perform some actions autonomously while remaining under human control when performing others. This is a consideration to which an accurate theory of accountability for autonomised harming will need to be sensitive.

4.2 The rectificatory imperative

Implicit in prominent discussions of 'responsibility gaps,' recall, is the Accountability Condition: the assumption that the possibility of accountability itself constitutes a condition for permissible deployment. A popular view is that AWS should be banned in part because the autonomisation of agency obscures traditional paths to responsibility. The standard position, among scholars and practitioners, is that human agents must remain in control over AWS precisely because AWS are not agents that can be held morally or legally accountable for their actions and decisions.[48]

Suppose that the possibility of accountability indeed constitutes a self-standing consideration relevant to determining the permissibility of building and employing autonomous systems with the capacity to harm humans. Indeed, there is intrinsic moral value in rectifying wrongs.[49] Our ability to address, mitigate, and meaningfully compensate rights transgressions is therefore worth protecting. Call this the *Rectificatory Imperative*. Insofar as the autonomisation of harmful actions will decrease the availability of certain rectificatory measures—say, as a result of moral responsibility disappearing from the picture—this may militate against autonomising actions that may transgress rights in a way that is impossible to adequately rectify *ex post*. Even if, as noted above, victims' claims to compensation persist in

[45] Emphasis added. Sparrow, 'Killer Robots,' 65.

[46] Andrew Smith, 'Franken-algorithms: the deadly consequences of unpredictable code,' *The Guardian*, 30 August 2018, https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger.

[47] Benjamin Wittes and Gabriella Bloom, *The Future of Violence: Robots and Germs, Hackers and Drones. Confronting A New Age of Threat* (Basic Books, 2015), 32.

[48] Peter Asaro, 'Autonomous Weapons and the Ethics of Artificial Intelligence' in Liao (ed), *Ethics of Artificial Intelligence*. See also Asaro, 'On Banning Autonomous Lethal Systems: Human Rights, Automation and the Dehumanizing of Lethal Decision- Making,' *International Review of the Red Cross* 94 (2012): 687– 709.

[49] I use 'wrongs' in a broad sense here, to include harms non-liable people might suffer as a result of justified rights infringements. Victims of all-things-considered permissible harms may still be wronged, insofar as their rights not to be harmed were, albeit justifiably, transgressed.

the absence of identifiable wrongdoers, a world in which duties of rectification are grounded in responsible agency is in better moral order than one in which remedial obligations are grounded in more morally removed features, such as capacity to compensate or group membership.

There are at least two objections to the Rectificatory Imperative as a ground for the Accountability Condition. First, the moral mechanism it presumes looks disconcertingly similar to that of paying indulgences—except that, rather than 'pay now to avoid punishment later,' it appears to advocate something along the lines of 'ensure the possibility of compensation later to permit causing harm now.' But this is not what the Rectificatory Imperative presumes. The view that there is independent value in rectifying wrongs need neither presuppose nor entail that the availability of compensatory measures *ex post* itself renders a rights-transgressing harm any more permissible *ex ante.* It just means that a world in which such wrongdoing can be meaningfully addressed is, all other things being equal, preferable to one in which rectificatory measures are unavailable.

The second concern has to do with the comparative value of rectificatory justice. This seems to me a more serious worry. How ought we to weigh the possibility of rectifying wrongful harms against other good outcomes, such as the possibility of reducing harms? If autonomous systems were likely to significantly reduce harms overall, should we insist that the importance of rectificatory justice outweighs that of reducing harm? That is, should we prefer a world in which more harms are caused, for which perpetrators can be punished and victims compensated, to a world in which less harm is caused, but without the possibility of accountability however defined? Call this the *Reduction/Rectification Conflict.* How we assess the *Reduction/Rectification Conflict* has significant implications. It may determine how AI's promise of reducing harm to innocent people is to be weighed against the value of being able to hold human agents accountable *ex post*.

The potentiality of rectificatory justice may, at least in principle, outweigh other considerations. After all, if all other things are equal, a world in which rectificatory measures are available is preferable to a world in which wrongful harms remain unaddressed. To the extent that certain rectificatory measures are available only for certain responsible agents' decisions and actions, the *ex post* possibility of restoring the moral equilibrium may therefore provide a *pro tanto* reason against delegating certain harmful actions to autonomous systems.

However, if other things are *not* equal, preferring a world in which harms caused are wrongful, rather than in the absence of moral agency, just so that rectificatory obligations may be more stably grounded *ex post,* would seem fetishistic.[50] Preferring a greater amount of compensable harm seems troubling not only because it would effectively view people as means to do the compensating or be compensated, in order to set the moral equilibrium right. It would also seem incoherent in putting the deontic nature of the harm—whether it is wrongful or detached from moral

[50] There may be complications regarding the number of people harmed vs the severity of harm, which I set aside here.

agency—before people's rights, since the former is determined by and matters precisely because of people's rights in the first place.

The *Reduction/Rectification Conflict* brings to the fore the question of how to weigh the value of minimising harms against the value of addressing harms wrongfully caused. This is not the place to resolve this issue, but here are three starting points. First, no matter how great the importance of rectifying rights transgressions, on any workable theory, the ultimate concern should be with preventing them in the first place. Second, even if, as per the Accountability Condition, the potentiality of rectification, broadly conceived, should constitute an independent consideration in determining the permissibility of autonomising harmful agency, how weighty this consideration is in comparison to others may well vary from case to case.

Third, here is another trade-off that might result from the possibility of harm reduction through the use of AI systems. The benefit of reducing harm, and saving lives, comes with unprecedented levels of control over who is harmed. In armed conflict, this may be a welcome development: if autonomous weapons technology allows for more accurate, individualised targeting, this may help direct risks away from the innocent. The possibility of autonomous driving in ordinary traffic, meanwhile, raises different questions. People who would have died in accidents caused by drunken, tired, or distracted drivers may, thanks to AVs, survive, whereas people who might otherwise have escaped crash sites as fortunate bystanders will be killed by AVs. One question this raises is whether—and, if so, how—decisions about whether to deploy AVs should account for the fact that the identities of those who will be harmed and those who will be spared by AVs, though fewer in number, will be different from those who would be harmed or spared by human agents.

5 Moral receptiveness revisited and perverting permissibility

Some philosophers have argued that one issue with delegating life-and-death decisions to robots is that they cannot replicate moral judgement. Duncan Purves, Ryan Jenkins, and BJ Strawser, for example, describe moral judgement as requiring 'either the ability to engage in wide reflective equilibrium, the ability to perceive certain facts *as* moral considerations, moral imagination, or the ability to have moral experiences with a particular phenomenological character.'[51] Since robots do not have these capacities, Purves, Jenkins, and Strawser argue, they cannot replicate moral judgement or make moral decisions 'for the right reasons.' And, they suggest, it is for this reason 'morally problematic' to deploy robots whose job it would be to make moral judgements.[52]

The notion of moral receptiveness outlined earlier helps explain *why* it might be morally problematic to task AI with moral decisions. What is more, lack of receptiveness to distinctly moral features of situations and actions might constitute a

[51] Duncan Purves, Ryan Jenkins, Bradley J. Strawser, 'Autonomous Machines, Moral Judgment, and Acting for the Right Reasons,' 852.

[52] Ibid.

pro tanto reason against autonomising harmful agency. What Purves, Jenkins, and Strawser do not make explicit is that just societies rely on the fundamental assumption that moral agents have the basic capacity to prefer right over wrong, and justice over injustice, with substantive conceptions of the good being subjects of reasonable disagreement. The very possibility of promoting or diminishing moral receptiveness, I submit, is itself duty-conferring when we face the option of bringing agents into the world who either possess or lack moral receptiveness. We have a basic duty to one another, to the extent that we can do so, to entrust only agents with moral judgements who possess the capacity to want to do the right thing.[53] We also owe it to one another to ensure that the people we bring into existence are capable of recognising distinctly moral features of situations and possess the capacity to relate to others as rights-respecting, duty-bearing agents.

Similarly to a duty to create just institutions, we have a duty to ensure that the agents we create are responsible ones. Delegating moral decisions to agents who, by nature, cannot want to do the right thing may accordingly constitute a distinct wrong. If AI lacks the capacity to look on persons as rights-bearing patients, this may militate against autonomising certain acts. It is for this same reason, I submit, that we would not appoint psychopaths as judges in criminal courts; why parents have a duty to bring up their children in a ways that enable reason, compassion, and responsible moral agency; and even why some societies limit home-schooling on distinctly democratic grounds. Members of just societies have a fundamental interest in other members functioning effectively as responsible moral agents. Insofar as autonomous systems lack this capacity, we have moral reasons not to delegate decisions about whether and whom to harm to them.

To be sure, not all human agents exhibit the sort of moral receptiveness I have sketched. But the fact that not all people think or act like sound moral agents tells us nothing about how we should think about the permissibility of delegating moral decisions to AI systems. It just means that the pool of agents to whom we would entrust high-stakes moral decisions is not coextensive with the pool of all humans.

5.1 Perverting permissibility

Here is a final implication, which we might call the *Problem of Perverting Permissibility*. *Perverting Permissibility* captures the peculiar possibility that autonomising harmful acts may render otherwise impermissible acts permissible. As indicated earlier, it is anything but clear how the difference between the stipulated permissibility of killing the one person to save the five in the standard trolley case and the impermissibility of killing the one to save the five in the footbridge variation could be translated into algorithmic codification, or be ‘learned’ through bottom-up approaches to machine learning. What renders killing the one morally impermissible

[53] How to assess this capacity is a different question which I will not address here.

in the footbridge version, we might think, is that there is something distinctly wrong with treating others as means, which is absent in the standard trolley case.[54]

Arguably, at least part of the wrong-making feature of treating others as means is agent-centred. It is not just worse to push the person off the bridge than it is to turn the trolley because it is worse *for that person* to be treated as a means than it is for the single person on the tracks to be killed as an unintended side effect of saving the five. And, considering the moral asymmetry between doing and allowing harm, all other things being equal, it is worse to do harm than to allow harm—not merely because it is worse for victims to have harm inflicted on them—say, by being hit on the head by a person—than to suffer an equivalent harm as a result of someone else's failing to prevent it—say, by being hit on the head by a falling branch that someone else could have caught. It is also worse because of what it says about the agent. Opportunistically inflicting harm on others is, all other things being equal, morally worse than failing to prevent others from suffering some equivalent harm.[55]

Another issue concerns the relevance of our intentions and, more broadly, the possibility of *justifiability.* Mainstream debates about the 'black box' nature of certain types of machine learning have focused on questions of interpretability and explainability of algorithmic decision-making and corresponding trust in AI systems.[56] One problem with 'black boxes' is that, without knowing *why* an AI system decided to act one way rather than another, it is not clear whether the harm it caused was *justified.* Philosophers disagree about the moral relevance of intentions, but all that matters for our purposes is the intuitive difference between the following set of cases:

> *Tactical Bombing:* A tactical bomber bombs an unjust enemy's weapons factory to destroy the enemy's weapons supply, but she foresees that this attack will also cause the deaths of some civilians nearby.
> *Terror Bombing:* A terror bomber deliberately attacks civilians to demoralise the enemy.

Now suppose that an AWS both kills a number of civilians but also destroys a weapons factory nearby. Without knowing *why* it did what it did, we will lack information necessary to morally assess the deaths it caused. If the AWS malfunctioned,

[54] After all, if the one somehow managed to run away before the trolley hit them, we would be all the happier. I will put potential complications, of which there are many, aside. For discussion, see Philippa Foot, 'The Problem of Abortion and the Doctrine of Double Effect,' *Oxford Review* 5 (1967); Judith Jarvis Thomson, 'Killing, Letting Die and the Trolley Problem,' *The Monist* 59 (1976): 204–217; and Thomson, 'The Trolley Problem,' *Yale Law Journal* 94 (1985): 1395– 415. For discussion of how uses of trolley cases (typically unwittingly) differ from their 'classical' application, see Frances Kamm, 'The Use and Abuse of the Trolley Problem: Self-Driving Cars, Innocent Threats, and the Distribution of Harm' in Liao (ed), *Ethics of Artificial Intelligence*. For a recent discussion of how the doing/allowing distinction might apply to autonomous machines, see Fiona Woollard, 'The New Trolley Problem: Driverless Cars and Deontological Distinctions,' *Journal of Applied Philosophy* 40 (2023): 49–64.

[55] I have jointly appealed to the DDA and DDE here, and there are numerous complications about their relationship, which need not concern us here. For discussion, see Warren Quinn, 'Actions, Intentions, and Consequences: The Doctrine of Double Effect,' *Philosophy & Public Affairs* 18 (1989): 334–351.

[56] For example, Zachary C. Lipton, 'The Mythos of Model Interpretability,' *Queue* 16 (2018): 31– 57.

or was hacked and made to harm innocent people, and the destruction of the weapons factory was a mere side effect, causing those deaths was wrong. If the AWS' objective was to destroy the weapons factory in pursuit of a just aim, and calculated that the unintended civilian deaths were an unavoidable and proportionate side effect, then the question is whether its presumptive lesser-evil justification was valid. One question is thus whether we can plausibly distinguish between intended harms and unintended side effects in the context of autonomised harming—not just because it is unclear to what extent non-human, artificial agents can be said to have intentions, though they might have 'objectives'; but also because their decisions might not be sufficiently transparent to be explainable, let alone to provide justifications.[57]

Assume that certain distinctions—for example, between doing and allowing – do not plausibly apply to AI systems, because artificial agents lack whatever grounds the moral significance of agent-centredness. Incidentally, even Asimov's Three Laws do not distinguish between doing and allowing harm. The first law says: 'A robot may not injure a human being or, through inaction, allow a human being to come to harm.'[58] Insofar as there are situations in which the wrong-making features of harming are agent-centred, and insofar as agent-centred considerations plausibly apply only to human moral agents, these wrong-making features are absent in autonomised agency. This generates the worry that harming in these cases might be permissible, given the absence of wrong-making, agent-centred features which would obtain only if the agent were a human moral agent. Thus the *Problem of Perverting Permissibility*: ordinarily impermissible acts might, by being autonomised, become permissible.

If pushing the person off the footbridge is impermissible for agent-centred rather than patient-centred reasons, and those wrong-making, agent-centred reasons are absent if the only agent on the footbridge is an AI-powered robot, then is it permissible for the robot to push the person off the bridge to save the five, although it would be impermissible for a human agent to do so? And, if so, does this render it permissible or impermissible to put the robot on the footbridge? On one hand—setting aside questions about what it means to say that something is 'permissible for' non-human, artificial agents—if it is permissible for the robot to push the person, then can it be impermissible to put it on the bridge? On the other hand, if pushing the person would be impermissible for a human agent, then how could it be permissible to autonomise the same act, and let a robot do it? This brings us back to peculiarities of a possible notion of 'autonomising' as a distinct mode of agency. So much for complications that arise from the possibility that involving AI might itself change the moral facts of the case.

One way forward here may be by focusing on agent-neutral, patient-centred considerations. For example, autonomous weapon systems might be programmed

[57] Greater explainability might come at the cost of decreased accuracy and, on some views, we ought sometimes to prioritise the latter. See Alex John London, 'Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability,' *Hastings Center Report* 49 (2019): 15– 21; see also Lipton, 'The Mythos of Model Interpretability.' One interesting question, which I will leave open, is whether a similar, somewhat paradoxical, trade-off might arise with regard to justifiability, where greater transparency, as a precondition for justifiability, comes at the cost of less accuracy with regard to moral judgements. That would be a vexing tension indeed.

[58] Isaac Asimov, 'Runaround' in *I, Robot* (Doubleday, 1950), 40.

according to a rule that says 'never kill civilians.' Thus focusing on reasons extrinsic to AI systems' acts, such as potential victims' rights not to be harmed, may help circumvent the *Problem of Perverting Permissibility*.

Here is a final thought. Even if there is a sense in which autonomised agency constitutes a *sui generis* area of moral activity, whether it is in this light permissible to autonomise—that is, delegate to AI—decisions about whether and whom to harm is a question about what we, as responsible moral agents, owe to one other. What this discussion showed is that this is in part a question about the permissibility of outsourcing rights-relevant decisions to tools that may not adequately respond to rights-based principles. The most vexing puzzles—such as the *Reduction/Rectification Conflict* and *Perverting Permissibility*—may thus arise not as a result of technological limits or gaps in accountability, but from the incompatibility of autonomised agency with essential features of rights-based views of what we owe to one another.

6 Conclusion

This paper examined little-appreciated considerations relevant to assessing the permissibility of developing and deploying AI systems with the capacity to inflict harms on humans. By articulating a series of problems that are in need of resolution, and for each starting to chart a path forward, this paper sought to illuminate issues that theories of autonomised harming will likely confront.

The paper first considered the challenge of how to integrate the notion of autonomised harming into our rights-based framework; it examined a set of considerations that might bear on the permissibility of employing autonomous systems, given the potentially limited degree to which human agent-centred moral principles retain their jurisdiction over autonomised agency. Accordingly, the first issue this paper identified was that of Agency Without Duties: the possibility that autonomous systems will act independently of human moral agents, yet without being constrained by moral duties in any conventional sense. The second issue the paper identified was that of Double Divergence: the possibility that the principles governing the permissibility of harming for AI systems differ from those governing the permissibility of harming for human persons in at least two ways. First, they might differ with regard to the *site* on which they apply: while the same principles govern the permissibility of harming for human persons in all contexts, different principles govern the permissibility of harming for AVs in domestic traffic on the one hand and AWS in armed conflict on the other. Second, these principles might diverge with regard to their *content:* principles governing the permissibility of harms caused by AVs might differ from those governing the permissibility of harms caused by human drivers in the same circumstances; and principles governing the permissibility of harms caused by AWS might differ from those governing the permissibility of harms caused by human combatants in armed conflict.

In addition, the paper identified two further, typically obscured considerations relevant to the permissibility of autonomising harmful acts: the fact that algorithms lack a necessary component of responsible moral agency, namely receptiveness to the distinctly moral features of situations; and the intrinsic value of rectificatory justice which may weigh against the possibility of reducing harms to innocent people. Both moral receptiveness and the Rectificatory Imperative, I suggested, may constitute distinct moral reasons against autonomising harmful agency, even if putting AI in charge promises to reduce harms.

Theories of autonomised harming must not merely account for the possibility that we cannot translate moral principles into algorithmic codification, replicate normative reasoning, that algorithms might make the wrong decisions, or the right decisions for the wrong reasons. They must in addition confront the possibility that autonomised agency presents a *sui generis* area of moral agency. Insofar as AI systems are at once autonomous and seemingly unconstrained by moral duties, certain properties pertaining to AI systems will be unique to autonomised agency. This creates the possibility of shirking the weight of moral decision-making and concomitant moral burdens to agents who cannot perceive them as such. The question before us remains what we owe to one another in light of these new technological possibilities.

In the end, the possibility of delegating moral decisions to AI confronts us with questions about key elements of our moral architecture: questions about which agents inhabit the moral system, over what types of agency the system presides, and what we think it means to interact with one another as rights-bearing moral agents as we increasingly relinquish our monopoly on moral decision-making.

Acknowledgements Versions of this paper were presented at the Institute for Ethics in AI at the University of Oxford (2023), the McCoy Center for Ethics in Society at Stanford University (2022), the International Association for Computing and Philosophy Meeting at Santa Clara University (2022), the Kennedy School of Government at Harvard University (2019), and the Amsterdam Graduate Conference in Political Theory (2019). I am grateful to audiences at each. Special thanks to Daniel Butt, Rob Reich, Mathias Risse, Milo Phillips-Brown, and Liesbeth Schoonheim for helpful discussion, and to Seth Lazar for his encouragement. Finally, many thanks to two anonymous reviewers for this Journal for their helpful comments.

Declarations

Conflict of interest The author declares that they have no Conflict of interest.