Philosophy & Technology (2021) 34:1013–1032
https://doi.org/10.1007/s13347-021-00448-5

RESEARCH ARTICLE

Liability for Robots: Sidestepping the Gaps

Bartek Chomanski[1]

Received: 12 October 2020 / Accepted: 14 March 2021 / Published online: 1 April 2021

Abstract
In this paper, I outline a proposal for assigning liability for autonomous machines modeled on the doctrine of *respondeat superior*. I argue that the machines' users' or designers' liability should be determined by the manner in which the machines are created, which, in turn, should be responsive to considerations of the machines' welfare interests. This approach has the twin virtues of promoting socially beneficial design of machines, and of taking their potential moral patiency seriously. I then argue for abandoning the retributive approach to machine crime in favor of prioritizing restitution. I argue that this shift better conforms to what justice demands when sophisticated artificial agents of uncertain moral status are concerned.

Keywords AI ethics · Autonomous artificial agents · Liability gap · Retribution gap · Robot ethics

1 Robots, Responsibility, and Retribution

What should the governance of autonomous machines look like? What policies should be instituted to regulate the production and sales of such machines? From a social welfare point of view, autonomous machines should be used in ways that, at a minimum, avoid unjustified harm to innocent people. This seems undemanding enough. It is probably a truism, however, that even the best-designed autonomous agents will not be perfect; therefore, they will sometimes cause unjustified harm. This is because, according to many writers on the issue, machine autonomy implies a degree of machine unpredictability. Since not everything the machines will do can be predicted, even by those with relevant expertise and information, some of their conduct could result in harm to innocent parties, despite the machine creators' best efforts. Unless the technology is banned from ever being created at all, then, we will

✉ Bartek Chomanski
b.chomanski@gmail.com

[1] Rotman Institute of Philosophy, Western University, 1151 Richmond Street North, London, ON, Canada

need a system for assigning legal (and perhaps moral) responsibility for machines' misdeeds. Call this the liability problem.

A different potentially salient side of the issue of creating autonomous machines is, to put it boldly, machine welfare. The more sophisticated the machines become, the stronger the case will be for taking into account their moral patiency. The question of whether there are morally weighty machine interests – interests that need to be considered in any moral calculus – will arise with more urgency than it is currently given.

Machine welfare is interestingly connected with machine responsibility – on a prominent line of thought (Danaher, 2016; Sparrow, 2007), for machines to be considered responsible for their acts, it must be possible to impose meaningful punishment upon them. Since punishment involves frustration of important interests (to punish someone, so the thought goes, is to inflict suffering on her), in order to be properly punished (and so, to be properly held responsible), the machines would have to have morally weighty interests in the first place.

Many would balk at the idea that even highly autonomous machines populating philosophers' imaginations will be able to suffer – which leads to a problem. Justice demands (so it seems) that wrongdoers be punished. If machines do wrong, but cannot suffer, then it's impossible for them to be punished for their wrongdoing. This, in turn, is an injustice. On the other hand, punishing someone else for what a highly autonomous machine does also doesn't strike many as just. Given these difficulties, one could make a general case against using highly autonomous machines in situations where they could conceivably cause harm – since when they do, by necessity, justice will not be served, whatever happens. This problem has been baptized the "retribution gap" (Danaher, 2016) in the literature, and I will adopt the label in what follows.

Reflection on the retribution gap demonstrates that matters of justice interact with metaphysics and philosophy of mind – but they also bring up obvious epistemic issues. Most importantly, it is not clear whether we will (ever[1]) be in a position to know whether highly autonomous machines have the requisite mental states that make punishing them and caring about them justified. It is possible, even likely, that the philosophical development of epistemic tools needed to make this determination will lag the technological developments leading to the creation of highly autonomous machines for whom the question of moral patiency may sensibly arise. Nonetheless, even in these circumstances, decisions will have to be made: will it be permissible (morally and legally) to build machines whose moral status is difficult to determine? If so, what sort of moral/legal status should they be granted?

My aim in this paper is to propose a solution to some of the problems raised by machine responsibility, machine welfare, and machine punishment. In particular, I will present a strategy for assigning liability for autonomous machines' acts and for holding machines criminally responsible, while uncertainty about their moral status persists. Parts of the solution have antecedents in the literature – I will content myself

[1] See Prinz (2003).

with adding more detail to the broad suggestions already existing in others' work, and with drawing connections that, to my knowledge, have not yet been explored.

In short, I will argue that we should link machine interests to machine responsibility and detach it from machine punishment.

2 Machines

What sorts of artificial beings am I discussing here? I want to cover a broad variety of futuristic machines whose possibility has been entertained by philosophers. I have in mind artificial agents that, *at a minimum*, are highly intelligent,[2] capable of unguided action, and moved by goals that are in principle susceptible to revision.[3] This characterization encompasses a number of different artificial agent types, from those capable "merely" of modifying their own states in certain ways to full-blown artificial persons in possession of genuine consciousness and volition.

Below is a sample of philosophers' characterizations of such machines:

- "While they will be programmed to make decisions according to certain rules, in important circumstances their actions will not be predictable. However this is not to say that they will be random either. … Instead the actions of these machines will be based on reasons, but these reasons will be responsive to the internal states — 'desires', 'beliefs' and 'values' — of the system itself. Moreover, these systems will have significant capacity to form and revise these beliefs themselves. They will even have the ability to learn from experience" (Sparrow, 2007)
- "a robot with the ability to make decisions on the basis of its own initiative(s) rather than preprogrammed commands, and of setting its own ends and/or achieving subordinate ends through means of its own devising" (Champagne & Tonkens, 2015)
- "[a] system that will gain knowledge and skills from its own behaviour, while learning from the features of the environment and from the living beings who inhabit it. This robot … will respond to stimuli by changing the values of its own properties or inner states and, furthermore, it will modify these states without external stimuli while improving the rules through which those very states change" (Pagallo, 2010)
- "[machines that are] artificially sentient and artificially intelligent. Such robots would not just seem to experience pain or pleasure, they would experience it; they would not just act like they have deeply held goals and values, but they would actually have them" (Petersen, 2017)

I will mostly use the term "machines" to cover the types of entities countenanced by the above authors. It is important to note that while the differences between the

[2] Approaching, reaching, or even surpassing humanlike intelligence.

[3] Exclusion of consciousness, by which I mean the capacity to instantiate phenomenal properties (or, undergoing states that instantiate phenomenal properties) is deliberate for reasons that should become clear in what follows. I want to discuss machines that may, for all we know, exhibit consciousness, without limiting myself to talking only about genuinely conscious machines.

machines described by, e.g. Petersen and those described by Sparrow are theoretically significant (Sparrow, and, following him, Tonkens & Champagne, deny sentience to their machines), it may be extremely difficult, if not impossible, to tell them apart in practice. It might be impossible to know, for instance, whether the programming and architecture sufficient for the emergence of the abilities that Sparrow's machines are supposed to have also gives rise to sentience and phenomenal consciousness. It is, I believe, an advantage of my view that the solutions to the liability problem and the retribution gap that I propose can go through regardless of settling this issue.

3 Principles of Machine Design

The claims I am defending in this section are, first, that machine design should be governed by the presumption that the machines have genuine welfare interests, even under conditions of uncertainty as to their real capacities and, second, that the way to apply these principles in practice is by adhering to the best-interest standard of surrogate decision-making.

To dramatize our topic, imagine an accomplished machine designer of the future, Ada, who decides to build a state-of-the-art highly autonomous agent, of the kind we've seen described above. Before getting to work, Ada must answer some questions, however, such as: What, if any, machine-centered[4] considerations should she be obligated[5] to follow in this enterprise? How should her design be constrained to accommodate the wellbeing of the machine she's building?

One obvious answer is that since the machine will not be conscious, there are no relevant machine-centered considerations (assuming that consciousness is necessary for having morally weighty welfare interests). Other values must therefore entirely determine the overall permissibility of constructing autonomous agents and guiding Ada's actions.

The problem with this suggestion is of course that it runs counter to the epistemic-technological mismatch highlighted above. As I mentioned, it may be extremely difficult to determine whether the machines we're dealing with are Sparrow's machines (human-level intelligent, reasons-responsive, but not phenomenally conscious) or (say) Petersen's machines (rational, conscious, sentient). So, decisions will have to be made before there's any certainty about the presence of consciousness in artificial agents. Under such uncertainty, treating the machines as mere automata could expose us to an unbearable moral risk: the moral risk of treating moral patients as if they were mere objects. On the other hand, presuming these machines to have genuine interests may hobble the designers with unnecessary burdens, thus slowing innovation and limiting the benefits that autonomous agents will surely bring society.

[4] I am assuming that there are obligations towards 'society at large' that Ada also incurs by creating such machines. For example, at a minimum, she should not build machines programmed to harm innocents. I briefly return to the question of how such obligations should best be enforced later in the article.

[5] I think here of legal rather than moral obligations. I leave it open whether Ada might have *moral* reasons to refrain from building certain types of legally permissible machines.

To resolve this tension, I'll argue that efficiency-related worries about having the question of design decided by machine welfare considerations – that is, by treating the newly created machines *as if* they had interests, regardless of whether they in fact do – are overblown. This is because, as we shall see, taking as-if interests into consideration still permits a broad variety of design approaches, so that many legitimate ends of the designers can be met. At the same time, it provides strong safeguards against mistreatment of the machines if it were to turn out that they do have morally weighty welfare interests themselves. Let us see then how operating on the supposition (a legal fiction, if you like) that machines have morally relevant interests would guide their design.[6]

Thankfully, we are not maneuvering in a vacuum here. There exist broadly recognized guidelines about how to proceed when making decisions for other entities with morally weighty interests[7]: the standards for surrogate decision-making familiar from the bioethics literature (Brock, 2014). These standards govern situations when a decision has to be made that concerns the treatment of another person, where that person is not in a position to give or withhold consent, so someone else – the surrogate – has to make the decision instead.

To simplify a little, the standards for the surrogate's decisions are as follows: *advance directives*, formulated by the patients themselves ahead of time, while still in full possession of their capacities; *substitute judgment*, where the decision is guided by what is known about the patient's preferences and values when the advance directive is unavailable, by trying to determine what the patient would have wanted; and the *best-interests* standard, where the other two methods are unavailable (this is the method used especially frequently in cases where the patient is a small child), and the decision is guided by what promotes the patient's interests best.

Since at the time Ada must make her design choices, the machine in question is not available to inform her considerations (by e.g. providing advance directives) we should adopt those surrogate decision-making principles which do not require antecedent access to another's preferences. That is, we should be guided by the best-interests considerations. Ada's decision about building an artificial agent ought, therefore, to be guided by what would be in the best interests of the machine she's building.

What interests should the artificial agents be imputed to have? In a sense, as it will turn out, the most important practical distinction will concern the scope of an artificial agent's *autonomy* – specifically, its capacity to determine its own goals and values. However, before we reach this step in the argument, let's consider how the application of the best-interests decision-making models can actually get us there.

[6] To reiterate: the claims that follow do not assume that machines do in fact have interests. If I simply assumed that they did, it would be uncontroversial that something like surrogate decision-making standards would have to guide their creation. My claim is that even if we don't know whether the machines qualify as moral patients, it is a good idea to produce them in accordance with these considerations.

[7] I have assumed that the mental capacities of the machines described in the quotations in Section 2 (other than consciousness) would be relevantly similar to those of an adult human being, rather than, say, a non-human animal. In my view, a (morally significant) being with the sorts of capacities described by Sparrow or Petersen would be disrespected if we were to treat it as we treat, e.g., pet dogs or lab mice. I will leave mostly to a side the question of how to deal with machines whose moral status is better compared to non-human animals.

First, following Degrazia (1995), I am assuming that the interests in question should be conceptualized by reference to wellbeing (Basl (2014) uses a similar strategy when discussing machine interests – I am essentially adopting his approach). The basic division among the theories of wellbeing is that between hedonistic, desire-satisfaction, and objective list theories. To simplify, hedonistic theories claim that wellbeing consists in the accumulation of pleasurable mental states; desire-satisfaction theories claim that wellbeing consists in getting what we want (or, what we would have wanted if fully informed and rational); objective-list theories claim that wellbeing requires the possession of certain objectively valuable goods, like friendship or freedom.

If promoting one's best interests means the promotion of one's wellbeing, then what one needs to do in order to promote the machines' best interests will depend on how one conceives of their wellbeing (or, as-if wellbeing).

In terms of constraints that adherence to the best-interests standards imposes upon the creator of the machine, both hedonistic and desire-satisfaction views of wellbeing permit a very broad degree of control over the machine's features. This is because, roughly speaking, the machine can be made to find 'pleasure' in – or desire – anything in the world, from world domination to serving a human being's every whim (see Petersen (2011, 2017) for some details on how this could work when designing intelligent, sentient robots). Consequently, Ada's freedom to decide her artificial agent's features will be greatest if she abides by either of these standards, since she can build the artificial agent to have any goals whatsoever (subject to some coherence constraints).[8]

A more objective understanding of the best-interests standard, such as that informed by objective list theories would impose stricter constraints on the creators of artificial agents. While deciding what should show up on the list is difficult (even for human wellbeing!), the main item I am proposing to consider here is autonomy, or, more specifically, an "open future." Open future is considered by some philosophers to be a component of best interests (Kopelman, 2007), and it can plausibly be thought to be a partial constituent of an autonomous agent's wellbeing.

To provide an agent with an open future means to refrain from foreclosing an excessive number of future options to her; that is, foreclosing now her ability to choose to pursue a diverse array of options at a later time, when the opportunity to choose manifests itself (Feinberg, 1980). Ada's decision in the present to build a reliably obedient machine, a machine desiring or finding great pleasure in obeying humans, would thus severely limit the machine's future options.

Using terms introduced by the philosophers quoted near the beginning of this paper, we can speak of machines designed with an open future as those who are rewarded (or at least not penalized) for altering their own hierarchy of goals or

[8] There might be complications when we consider more sophisticated desire-satisfaction theories, focusing on idealized or informed, rather than actual, desires. But it seems that since we do not have access to the machine's preferences prior to creating those preferences, we cannot infer its idealized preferences either.

"values." Those would be the machines for whom the determination of their highest goal – what they most "desire" – is not entirely up to their designers' decisions, and their goals need not be kept stable and unchanging.[9]

This also helps explain why adopting the objective list theory for our purposes gives rise to the focus on the open future,[10] rather than other components of objectively good life. On most objective list views, other objective goods on the list include achievements, knowledge, and meaningful relationships. Suppose Ada is a pluralist about the objective goods on the list. That is, she thinks there is no single underlying principle, no common nature explaining that in virtue of which they are all good. Suppose also that these values are incommensurable (see Mason (2018, especially Section 4)). As a result, it might be impossible for Ada to antecedently create a metric on which all these goods could be arranged (either because incommensurability implies incomparability, or because to choose between incommensurable values one requires practical wisdom, which is a quality developed through exercise of choice over a person's life, rather than something that can be programmed from scratch). Hence, Ada cannot impose a rigid hierarchy on the machine's goals, with some single goal as its highest one. Without imposing such a hierarchy, Ada cannot determine what the machine will most desire and how it will handle the inevitable tradeoffs between different goods. Consequently, Ada will have to leave this sort of determination to the machine itself as it makes choices over its lifetime, hence, open future. If, in contrast, Ada were a monist about the objective good, it seems that the design procedure for building a machine to pursue that good would be no different than on the desire-satisfaction model (build the machine to desire the single "super"good the most).

We could think of the machines' being able to choose from among a variety of goals as permitting, or perhaps rewarding self-generated alterations to their utility function. While it's controversial whether AI systems would generally be likely to change their utility functions (Omohundro, 2008), one can nevertheless imagine not-unusual circumstances where rational agents are able and willing to do so (Miller et al., 2020;

[9] Lest this be construed as contradicting the description of machines given by the quotations at the beginning of this paper, we may think of machines built in accordance with the desire-satisfaction or hedonistic models as having their highest value ('serve humans', 'be a nanny') set by the designers, while all the subordinate goals and ways of achieving them could still be formed and revised by the machine itself.

[10] There are other reasons for prioritizing open future. For one thing, if one agreed with Mill (1859/2015) that certain valuable capacities, like "perception and judgment" can only be cultivated through autonomous choice, and if one further thought that such capacities are required for the pursuit of other goods on the list (like knowledge, achievement, and friendship), one could use this reasoning to prioritize open future. More pragmatically, it also seems that, out of all the goods on the list, it is only the open future that could potentially impact our judgments regarding the allocation of liability.

Totschnig, 2020). Machines designed to have an open future would thus have ample opportunities to modify their utility functions as they gather experience.[11]

To sum up: reflection on surrogate decision making, as applied to potentially morally significant artificial agents, ultimately reveals two broad paradigms of agent design that emerge from adhering to the best-interests criterion. On one view, taking machine welfare into account means allowing the agents to make choices about their own (highest, most valuable) goals, and to revise and modify them; on the other view, such autonomy is not necessary for machine wellbeing; what matters is that machines are able to achieve their goals (get what they "want" or maximize "pleasure").

What would follow practically from these considerations for the creators of autonomous agents? On a more dogmatic view, Ada should be legally obligated to build her machine in a way that conforms to *the correct* account of best interests and wellbeing (the account that's correct for machines, that is, which might be different from the account that is correct for human beings). This, to put it mildly, is a tall order, difficult to square with commonly accepted normative and epistemic limitations of the state. However, even without settling the issue of which theory of wellbeing we should favor, it might still be possible for a policy to be successfully guided by the preceding considerations. In particular, I will argue that while the designers should be free to choose which standard of machine wellbeing they abide by, that choice will determine the level of their control over the machines' features, and consequently, the level of liability they will bear for what their machines do.

4 Wellbeing, Control, and Liability

Consider the following principle: the greater the control one has over their autonomous machines, the greater responsibility they bear for the machines' actions.[12] Call this the control-responsibility link, or CRL. Control over machines can be manifested in two ways: designers control the machines' conduct through programming (in our simplifying abstraction, by setting the machines' goals); users control the machines' conduct through issuing commands to the machines to carry out tasks.

In the first instance, control always lies with the designer. It is only once the machine is sold to, or hired by, someone else that the user acquires control over its conduct.

[11] Though she doesn't express it in exactly those terms, I think that Rini (2017) proposal about how to "raise" an autonomous agent – i.e. in the way analogous to how we provide moral education to children – could serve as an apt example of creating an open-future machine.

[12] As will become clearer in the remainder of this paper, my idea of control differs from the concept of "meaningful human control" discussed primarily, though not exclusively, in the context of military robots, most famously by Santoni de Sio and van den Hoven (2018). I do think that my points are compatible with their account, however. CRL is also similar to one of EU's Expert Group's (2019) guiding principles used to determine who counts as the operator of the machine for the purposes of liability allocation.

Combined with the surrogate decision-making models just discussed, CRL allows for differential responsibility assignments depending on which decision-making standard machine creators choose to follow. Those who elect to abide by a more permissive model, such as the desire-satisfaction model, would bear correspondingly more responsibility for their creations' acts than those who abide by the stricter open future model.

Let us leave the question of moral responsibility for a different discussion. Criminal responsibility will be addressed at a later stage in this paper. Now I want to focus on civil/tort liability. In this restricted context, CRL dictates that those possessing greater control over their machines will bear correspondingly greater liability for what their autonomous machines would do.

There is no shortage of proposals for handling the problem of liability for machines' acts (Asaro, 2012; Chopra & White, 2011; Diamantis, 2020a, b; Gunkel, 2020; Nyholm, 2020; Pagallo, 2010; Schaerer et al., 2009; Turner, 2019). I will develop in a little more detail the suggestion entertained by Diamantis (whose solutions are closest to my own, though his focus is on currently-existing algorithmic entities), Turner, Chopra & White, and Asaro for using the doctrine of *respondeat superior* to think about creators' and users' responsibility for the harms that autonomous machines will cause. The *respondeat superior* doctrine has a rich history in common law and is currently used in various jurisdictions to allocate liability to employers for tortious conduct of their employees – specifically the kind of conduct that falls within "the scope of employment."[13] Different jurisdictions have differed over the years in how broadly "scope" is interpreted, variously expanding and limiting it (for more detail see Young (1990)).

I propose (along the lines suggested by Asaro, Chopra & White, and Turner) to treat the creators and/or users of highly autonomous machines as employers, and the machines themselves as employees, *for the purposes of assigning liability for the machines' tortious conduct*. As a result, what acts are covered by a particular liability assignment will depend on which best-interests standard is adhered to by the creator, in accordance with the CRL: the greater the control over the machine's features (including, crucially, its hierarchy of goals), the broader the "scope of employment" – that is, the scope of the machine's activities that, if harmful, will generate liability for the user and/or designer, because they can be assumed to fall under some measure of control by either party, or because they are carried out for, or on behalf of, the user.

For the purposes of *respondeat superior*, an employee's act counts as within the scope of employment "[i]f an employee commits a tort while performing work assigned by the employer or while acting within a course of conduct subject *to the*

[13] As the *Restatement (Third) of Agency* puts it, *respondeat superior* essentially means that "An employer is subject to liability for torts committed by employees while acting within the scope of their employment" (§2.04).

employer's control,[14] … unless the employee was engaged in an independent course of conduct *not intended to further any purpose of the employer* [emphasis added]" (Restatement (Third) of Agency (§7) 2006). In light of this, it strikes me as attractive to designate *everything the machine does* as being within the scope of its "employment" if *the machine is designed in adherence to the desire-satisfaction or hedonistic* understanding of best interests.

These machines, created without the "open future" requirement, would presumably be built to reliably obey their users' wishes and desires (indeed, this scenario – creating intelligent machines to be humans' obedient servants by making them enjoy or desire such conditions – is specifically raised both by those who worry about machine servitude Bloom & Harris, 2018; Musiał, 2017; Walker, 2006), and by those who see it as morally permissible (Petersen, 2011)). Generally speaking, a reliably obedient robotic servant, one that doesn't set its own highest goals, would at all times be under someone's control. Since the machine's purposes (or at least its highest goals) would have been determined by the designer, one can think that it always remains under the designer's control. On the other hand, once the machine is assigned tasks to carry out, it would be the assigner (user) that assumes control. Thus, either the designer or the user would be held liable for the machine's conduct, whatever it does. In the next section, I will discuss how exactly the apportioning of liability between designer and user could proceed.

This extremely broad understanding of *respondeat superior* is redolent of the ancient origins and medieval application of the doctrine, when masters bore "full responsibility" for their domestic servants' acts (Wigmore, 1894), and conforms to Bryson's (2010) dictum that machines "are entirely our responsibility. We determine their goals and behaviour, either directly or indirectly through specifying their intelligence, or even more indirectly by specifying how they acquire their own intelligence." Nevertheless, my proposal differs from Bryson's in that it allows for different rules for differently designed artificial agents, and it does not demand they lack moral patiency.

This way of assigning liability would prevent the production of machines that enjoy or desire performing clearly immoral acts, such as violating rights and causing unnecessary pain and suffering. Being held liable for injuries of this sort would very likely discourage designers from the deliberate or reckless construction of immoral machines – and would incentivize care in equipping the machines with constraints on their behavior – even in the absence of specific regulations concerning the moral code to be prescribed to (or inscribed

[14] What exactly *is* control? I find it useful to supplement the common law doctrine of respondeat superior with the way in which the courts have interpreted the statutory provisions regarding "controlling persons" found in securities law – therein, "[t]he term 'control' means the possession, direct or indirect, of the power to direct or cause the direction of the management and policies of a person" (C.F.R. 17 § 230.405(f)). As Kuehnle (1988) explains, "[t]he definition of control is recognized generally as embracing indirect means of influence. Even courts requiring culpable participation have recognized the regulatory definition of control and its relatively great breadth. This broad definition of control appears to apply to the wide scope of possible forms of control." Applied to our purposes, this means that not just direct programming of machines to do something would count as controlling them, again hewing closely to Bryson's quotation above.

in) machines (analogously, it would also disincentivize the issuing of immoral or careless commands by the users). A machine likely to violate rights (or even to pay insufficient regard to them) would be too costly to maintain in virtue of exposing the user to a high risk of expensive lawsuits.

Consider now the other possibility – adhering to the "open future" conception of best interests when building a highly autonomous machine. In cases such as these, upon their launch, the machines are provided with a sufficiently broad range of meaningful options: they have the capacity and the opportunity to select the preferred one autonomously – and their hierarchy of goals is malleable, with a broad selection of life plans available to them. Machines of this sort would be guided in their decision-making by more than their designers' vision of what the machines should want and what sorts of purposes they should serve. The creators' control over such machines, then, is more limited than in the previous case. By CRL, so should be their liability.

Where machines of this sort are in the service of another person, individual or corporate, it seems that whatever law governs human employer liability could be directly applied to liability for machines; thus, for example, if ordinary *respondeat superior* were the dominant doctrine, the same set of rules regarding the scope and course of employment would apply to machine and human employees alike. Consequently, it would be possible for machines thus designed to have a sphere of actions for which *they alone* (and not their creators or consumers of their services) bear responsibility, when whatever it is they do falls outside the scope of their employment (I will later revisit the question of how to hold such machines responsible and how to extract compensation from them for harms they cause outside of "work"[15]).

Machine creators would still be incentivized to be careful in building in constraints on machines' behavior even in this case. Employers would, after all, be unwilling to hire someone desiring to cause unnecessary harm to others (or even someone careless about causing such harm) – as they would be liable for their misconduct under ordinary *respondeat superior*. Thus, it would not be commercially viable to produce machines with a warped or hastily put together moral code. It is likely that employers (potential users) would be loath to take on a risk of enlisting such machines in their service even if they wouldn't be responsible for *everything* the machines do.

Consequently, there would be demand-side pressures to build decently behaving machines under the system of ordinary *respondeat superior*. Thus, even machines with an open future would likely be equipped with a socially beneficial moral code.

[15] Determining what acts count as outside the scope of employment might become a difficult problem, but it is no different than making the same determination for human employees, and there is a significant amount of case law where the courts wrestle with this. It would not be a new issue generated by autonomous machines.

5 Whose Responsibility?

Who should be held responsible for the machine's actions? Starting with machines designed in accordance with the desire-satisfaction/hedonistic models of best interests it would be that person or organization whose relationship with the machine is closest to employment or (if the machines lack moral patiency) ownership. In the first instance, that would be the creator, who, in virtue of determining the machine's main purposes, would be exercising control over its actions. However, once such a reliably obedient machine is sold to/hired by another party, it is the other party that assumes control – that tells the machine what to do – and derives benefits from the machine's services. When control transfers, so does liability.

Companies and individuals could be allowed to reallocate liability via contracts, either between buyers and sellers or with third parties (e.g. insurers). Where liability is entirely transferred to the user, machines would probably be less expensive than in cases where liability wholly or partially remains with the creator. What should remain constant, though, is that *every* act of the machine that falls within the scope of its employment will generate liability for *some* third party.

One may worry that this arrangement would fail to protect users from design flaws in the machines. E.g. suppose I believe, in accordance with RoboCorp's description, that the robot they built that I am using to perform security duties around my mall is programmed to refrain from bodily contact with humans. One day, the robot causes an injury to an innocent third-party by hitting them in the face. While, under the proposed system, the liability for the injury lies with the user of the robot, it seems that, regardless of the contract between me and the manufacturer, RoboCorp should bear at least some, if not all, responsibility for what happens – after all, the robot seems to have acted in a way that RoboCorp explicitly said it would not act – at the minimum, the user should have the opportunity to seek recourse in circumstances of this sort.

Difficult cases like that, where it's unclear whether the injury is due to the manufacturer's error in programming might still be possible to litigate. Victims could be allowed, for instance, to sue manufacturers for an analog of "negligent training" (Fenton et al., 1991; Moore, 1988). As Moore (1988) explains, in lawsuits involving injury sustained at the hands of private security personnel, "[American] courts have made it clear that they will consider a lack of training a basis for a cause of action or support for a cause of action." A similar approach could apply to robotic workers. When the machines act in ways contrary to manufacturers' specifications, or in ways that manifest insufficient care in programming the machine, e.g. when a security bot built to restrain but not injure ends up causing bodily damage, it might be possible for the victims of the machine to seek damages from its manufacturers for negligent training – that is, for a failure to properly program the machine (as happens with such suits under current law, they could be brought concurrently with suits against the employer). Manufacturers would be negligent in training their robots in cases where the robots are unprepared for the environment in which they are

employed, as long as there is an understanding, implicit or explicit, that the manufacturer advertises their machines as able to perform up to a certain standard in such environments.[16]

Similarly, it's possible that the courts will consider it negligent to build an "open future" machine with a moral code that allows it to cause harm (violate rights, inflict unnecessary pain) when acting outside its scope of employment. If so, then it might be possible to sue the manufacturer of an "open future" robot even when the harm the robot causes is done without external influence. Perhaps only sufficiently serious harms caused by machines could be actionable in this way. The same approach, relying on negligent training, could be adopted in some cases where a machine harms its own user – again, though, the exact contours of how and when this kind of harm would be actionable need not be antecedently decided.

If this line of thinking is accepted, then the duty to train would always remain with the manufacturers of autonomous agents, even if breaching it only serves as the basis for a claim in cases of serious injury or death. Once again, this could incentivize the careful construction of the machine's moral constraints, for those machines that will have the potential to inflict serious harm on others.

Now, this may raise some issues: first, it could be considered an affront to the machines' autonomy, if it is of a piece with a typical human being's autonomy, to never be held *fully* responsible for one's actions; second, given that someone else pays for the machine's misdeeds, the machines themselves could be incentivized to engage in reckless or harmful behavior more often than socially optimal.

In response to the first problem, I would argue that, first, it's possible that *not all* of the machine's acts generate liability for its designers – only those that cause serious enough harm; second, the "tethering" (to use Johnson & Miller's (2008) term) of machines to their creators reflects the intuitions behind a pair of ideas: one, the already mentioned control-responsibility link, and two, the contrast between building machines and raising children (see e.g. Grodzinsky et al. (2008); Schulzke (2013); Schwitzgebel & Garza (2015); White & Baum (2017)). No single individual or organization (not parents, not schools) has as much influence (individually or collectively) over children's character, preferences, and values, as our hypothesized machine creators have over the analogous features of their machines. For this reason, the latter's responsibility for how their creations "turn out" and what they do must be greater than the former's. That this idea should

[16] One could also imagine a system where the user/employer is liable on the respondeat superior basis, but may be allowed to recoup their losses from the manufacturer if they can demonstrate a fault with the machine's training/design. Indeed, as Barfield and Pagallo (2020) point out, "within the broad umbrellas of tort … law, there are multiple specific (*and often simultaneous*) theories of liability that can be asserted … All of these liability theories can arise in the context of AI." My proposal is consistent with such a multipronged approach.

find its practical expression in different approaches to liability assignments for machines and children/adults is thus no surprise.

6 Machine Liability

In answer to the second problem, one may suggest that machines could themselves be held liable for their misconduct (just as *respondeat superior* permits suits against employees as well). To the extent that they can hold assets, this is not impracticable even for machines that do not ascend to the heights of moral patiency (see Pagallo (2010) and Katz (2008) on applying the Roman law idea of *peculium* to machine liability). On the other hand, given one of the main justifications for *respondeat superior* – that compensation be sought from those who are better able to spread its cost – whether the machine or its principal (or both) are sued may depend on who is in a better position to spread the costs of liability.[17]

Indeed, it might be possible for a widespread system of liability insurance schemes to arise, with insurance companies taking on a significant role in determining the way in which the machines are designed,[18] in exchange for assuming liability when harm is caused. This, of course, could be applied both to liability arising through *respondeat superior* as well as that arising out of negligent training. In any event, whoever ends up on the hook for the machine's conduct would be incentivized to minimize its propensity for harmful behavior.

Given that being held civilly liable for a wrong does not necessitate the defendant being punished, there is no problem of justice not being served under any of these schemes, whatever the true capacities of the machines turn out to be. Justice does not require punishments for tortious conduct.

7 Punishing Machines

If the above account is on the right tracks, then the questions of machine liability can be handled by adopting and/or adapting existing legal rules, even when machines acquire an unprecedented degree of autonomy. However, it might seem that in some cases justice demands more than the payment of damages. In some cases, it seems, justice demands that the guilty be punished. The difficulty of applying this seemingly obvious approach to machines (even if we assume that the machines under discussion are capable of having a guilty mind, a necessary condition for most kinds of criminal

[17] There are also practical problems about whether all robots should be provided with such assets, how much could each robot be allowed to hold and whether there should be a ceiling on the amount it could be sued for. I see no in principle reason, however, why such issues couldn't be solved by legislation, legal precedent, or some other means.

[18] Since insurers would have an interest in machines not engaging in tortious conduct, they would be incentivized to insist that the machines be equipped with a robust capacity for moral reasoning.

conduct) is the already mentioned retribution gap. To repeat: just punishment implies suffering, and machines cannot suffer, so they cannot be justly punished. It is also generally unjust to *punish* a third party for a crime they did not commit. So, whatever we do, justice will not be served.

Even if a decision is made to treat machines as moral patients – as if they have genuine morally weighty interests that can be frustrated, one could still expect public skepticism as to whether justice is really done when the machine is punished. As Sparrow (2007) notes:

> In order for a machine to be capable of being punished … it must be possible for it to be said to suffer. Furthermore, *its suffering must be of the sort that we find morally compelling*. … In order for our treatment of the machine to count as punishment, it must be capable of *suffering in ways that might motivate the same set of responses that we have as a matter of course to human beings*. It must be such that we could understand someone saying that they felt sympathy for it, or grief, or remorse, if this suffering turned out to be unnecessary. [emphasis added]

Given how demanding a standard this is, even if machines are "punished" in some way, it could be that, in the eyes of many, justice will not have been served, as their suffering will not be seen as "morally compelling," and observers won't feel grief or sympathy for machines who turn out to "suffer" undeservedly.

However, as Danaher (2016) points out, retribution is not the only way to handle criminal cases involving machines. The problems with applying the retributivist paradigm that Sparrow assumes in his famous article could well propel us towards exploring alternatives to punitive measures. One such alternative is the *restitutive* approach to criminal justice, outlined in much detail by, e.g., Randy Barnett (1977). This approach is distinguished from the standard retributive paradigm by focusing on compensating the victim – inasmuch as possible – rather than on punishing the perpetrator. In Barnett's words, "[t]he idea of restitution is actually quite simple. It views crime as an offense by one individual against the rights of another. The victim has suffered a loss. Justice consists of the culpable offender making good the loss he has caused. It calls for a complete refocusing of our image of crime."

There are no conceptual difficulties in ensuring that machines provide appropriate *restitution* – whether the victim is compensated in no way depends on the perpetrator possessing some relevant mental state. It depends, simply, on furnishing the required compensation. This can be done by machines themselves, or by whoever is liable for their misdeeds. As Barnett points out, under the restitutive paradigm, the distinction between civil and criminal law "for most purposes … collapses." If the "purposes" include vicarious liability, then there is no problem in applying the vicarious liability assignments I discussed above to machine crime.[19]

Although vicarious liability for crimes – including homicide – seems like an alien concept, it is by no means a novel one – particularly when punishment is replaced with compensation. The medieval institution of *wergild* (compensation paid for injuring

[19] Indeed, if this claim is accepted, then we should no longer speak of machine *crime*. *All* harmful acts such machines commit would better be thought of as torts.

or killing another person), for example, allowed for (or even required) the *family* of the offender to pay the compensation to the victim or the victim's family, either wholly or partially (Brown, 2011; MacCormack, 1974; Phillpotts, 1974). Incidentally, this does not seem to have led to violence and chaos (Benson, 1990; Davies, 1969). Indeed, some have taken the wergild system to be more just than the currently reigning paradigm (Friedman, 1989). Thus, my solution is not unprecedented. If one takes Sparrow's and Danaher's worries seriously, it also seems fairer than the alternative of imposing meaningless punishment on machines *without* compensating victims (or punishing third parties who lacked the *mens rea*).

Moreover, the restitutive paradigm would probably be easier to accept by the general public in cases where doubts remain as to the appropriateness or meaningfulness of punishing the wrongdoers – consequently, it looks tailor-made for the purpose of handling criminal acts committed by machines. For comparison, imagine a human criminal who is generally believed to suffer no setbacks to his interests by being imprisoned. Perhaps it is believed that, due to an unusual mixture of psychological and sociological factors, he enjoys his incarceration and neither his health nor his financial or personal situation will deteriorate as a result of it. In such circumstances, one could easily think that justice would be better served if the criminal were made to compensate his victim instead of going to prison – my guess is it would also be a popular sentiment among the general public. Indeed, I think reflection on this case may lead us to a more general principle that we can apply to the problem of machine culpability, namely: *if punishment is unlikely to play its retributive role, then the justice system should prioritize making the victim whole*. This principle can be endorsed by retributivists and their restitutionist detractors alike and thus evades the controversy attendant on replacing retribution with restitution wholesale.

As another comparison, consider the mostly medieval practice of holding animals criminally responsible for the harms they cause (Hyde (1916) describes many such grotesque cases). Imagine if society settled on a system in which farm animals are held criminally responsible for injuring or even killing a person while the victim (or their family) receives no compensation for the sustained injuries.[20]

It is, I think, clear that it would be an improvement over the system just described if the animal *owners* or caretakers were to bear civil liability for their animals' misdeeds, and animals themselves would not have to face criminal sanction. Why we should prefer the latter is well-explained by the principle articulated above – punishment of animals in these cases is unlikely to play its retributive role, hence it is better – more just – to prioritize compensating the victim instead.

Except in special circumstances, where, e.g., it can be proven that the machine was programmed or explicitly instructed by someone else to commit a crime (or to achieve its goal no matter what, with disregard for others' rights and suffering), no criminal responsibility need to be imputed to the machine's principals. In exceptional cases, they can be treated as, e.g., accomplices or conspirators.

[20] Interestingly, just such a rule is mentioned in the *Exodus*: "If an ox gores a man or a woman to death, then the ox shall surely be stoned, and its flesh shall not be eaten; but the owner of the ox shall be acquitted." (21: 28, New King James Version).

This approach circumnavigates the "retribution gap" and still allows for at least some justice to be done to victims of machine crime, despite whatever doubts one may harbor about the machines' susceptibility to punishment.

8 Conclusion

There are a few more virtues of the system I am proposing.

First, it would help address many problems of machine liability without the need to "solve" machine consciousness, thus evading the technology-epistemology mismatch. To implement it, we would not even need to answer the question of the machines' moral status.

Whether we decide that sophisticated artificial agents showing outward signs of consciousness and intelligence are to be treated as persons, objects, or are to enjoy some intermediate status – all that I have said so far stands. Just as a third party can be held responsible for harm caused by an object, they can also be held responsible for harm caused by a person. Other than by perhaps altering the details of individual cases, the decision to grant machines moral status would not change the broad-strokes picture of liability assignments I am painting here.

For comparison, it appears that there is no reason to think our practices for assigning liability to animal owners should have to change if the infamous Cartesian[21] position on animal pain (namely that they don't feel any) turned out to be correct. What matters for animal owners' liability are physical, public acts done by animals, rather than their internal states. Whether we think they can suffer is irrelevant. Analogously, machine phenomenology would not matter for liability assignments.

Secondly, the system I am outlining allows the creators of machines a significant degree of autonomy when constructing their products. They would not be required to follow a single paradigm but could instead rely on their own judgment as to what type of machines it's best to make.

It also lets us lay to rest the worries raised by thinkers such as the already mentioned Bryson whose opposition to treating autonomous machines as rights-bearers springs (at least in part) from the idea that such an approach will allow machine creators to evade responsibility for wrongdoing (see also Bryson et al. (2017)). As I have argued, my proposal is consistent with treating machines as rights-bearers and, at the same time, it *burdens machine creators and/or users with enforceable duties regarding the machines' harmful actions*. Just as it's possible to have employers held responsible for their rights-bearing human employees' conduct, it is possible to give machines rights *and* hold their "employers" or designers responsible for the machines' wrongdoing.

In presenting my argument, I have avoided criticizing alternative approaches to machine liability. This is because I do not envisage my proposal to be mutually exclusive with those strategies. As has been recognized elsewhere (Expert Group on Liability & New Technologies, 2019), a one-size-fits-all system of rules is unlikely

[21] While Descartes is commonly read as denying that animals feel pain, see Cottingham (1978) for a critique of this interpretation.

to be able to handle all the problems arising out of the widespread adoption of autonomous artificial agents. I am merely outlining the virtues of my proposal, especially when it comes to machines for whom the question of consciousness can arise. I am not arguing that it replace all other liability assignments.[22]

On the whole, the paper shows that it is possible to create a strategy for handling questions of liability for autonomous machines that permits a broad variety of legitimate designer aims to be pursued, and that can be called upon to protect the interests of machines if any such there be, all the while doing so without prejudging important – perhaps intractable – philosophical controversies.[23]

Availability of Data and Material n/a.

Code Availability n/a.

Declarations

Conflicts of Interest/Competing Interests n/a.

References

American Law Institute. (2006). *Restatement (third) of agency*. American Law Institute Publishers.

Asaro, P. (2012). A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot Ethics* (pp. 169–186). MIT Press.

Barfield, W., & Pagallo, U. (2020). *Advanced Introduction to Law and Artificial Intelligence*. Edward Elgar Publishing.

Barnett, R. E. (1977). Restitution: A new paradigm of criminal justice. *Ethics, 87*(4), 279–301.

Basl, J. (2014). Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines. *Philosophy & Technology, 27*(1), 79–96.

Benson, B. L. (1990).The enterprise of law : justice without the statePacific Research Institute for Public Policy

Bloom, P., & Harris, S. (2018). It's Westworld. What's Wrong With Cruelty to Robots? *The New York Times*.

Brock, D. W. (2014). Surrogate Decision Making. In B. Jennings (Ed.), *Bioethics* (4th ed., Vol. 6, pp. 3037–3040). Macmillan Reference USA.

Brown, W. C. (2011). *Violence in Medieval Europe*. Pearson.

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close Engagements with Artificial Companions* (pp. 63–74). John Benjamins.

[22] This is not to say, however, that my approach cannot be adopted in less futuristic scenarios. Consider, for instance, a hypothetical trading algorithm that engages in insider trading or price collusion while deployed by a corporation. Or consider the problem of responsibility for harms caused by autonomous vehicles. While I do not have the space to develop these suggestions here, I think the respondeat-superior based system of liability allocation could work here as well. In the former case, under respondeat superior, the corporate user of the price-setting algorithms would be held liable for these sorts of activities. In the case of AVs, we could have a system where either the designer of the car or its user (perhaps under the doctrine of "borrowed servant") would be held liable for the crashes. I leave the details to be developed in future work.

[23] I am grateful to Dan Lizotte, Anthony Skelton, and two anonymous reviewers for this journal for invaluable comments and suggestions on a previous version of this manuscript.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. Artificial Intelligence and Law, 25(3), 273-291. https://doi.org/10.1007/s10506-017-9214-9.

Champagne, M., & Tonkens, R. (2015). Bridging the responsibility gap in automated warfare. *Philosophy & Technology, 28*(1), 125–137.

Chopra, S., & White, L. F. (2011). *A legal theory for autonomous artificial agents*. University of Michigan Press.

Cottingham, J. (1978). 'A Brute to the Brutes?': Descartes' Treatment of Animals. *Philosophy, 53*(206), 551–559.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299–309.

Davies, R. R. (1969). The Survivial of the Bloodfeud in Medieval Wales. *History, 54*(182), 338–357. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-229X.1969.tb02328.x.

Degrazia, D. (1995). Value theory and the best interests standard. *Bioethics, 9*(1), 50–61.

Diamantis, M. (2020a). The extended corporate mind: When corporations use AI to break the law. *North Carolina Law Review, 98*, 893–932.

Diamantis, M. (2020b). Algorithms acting badly: A solution from corporate law. *George Washington Law Review* 89: np.

Expert Group on Liability and New Technologies. (2019). *Liability for artificial intelligence and other emerging technologies*. Retrieved from https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608. Accessed 8 Jan 2021.

Feinberg, J. (1980). The Child's Right to an Open Future. In W. Aiken & H. LaFollette (Eds.), *Whose Child? Children's Rights, Parental Authority, and State Power* (pp. 124–153). Rowman and Littlefield.

Fenton, J. W., Ruud, W. N., & Kimbell, J. A. (1991). Negligent Training Suits: A Recent Entry Into the Corporate Employment Negligence Arena. *Labor Law Journal, 42*(6), 351.

Friedman, D. D. (1989). *The machinery of freedom : guide to a radical capitalism* (2nd ed.). La Salle, Ill.: Open Court.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology, 10*(2–3), 115–121.

Gunkel, D. (2020). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology, 22*, 307–320. https://doi.org/10.1007/s10676-017-9428-2.

Hyde, W. W. (1916). The prosecution and punishment of animals and lifeless things in the middle ages and modern times. *University of Pennsylvania Law Review and American Law Register, 64*(7), 696–730.

Johnson, D. G., & Miller, K. W. (2008). Un-making artificial moral agents. *Ethics and Information Technology, 10*(2), 123–133. https://doi.org/10.1007/s10676-008-9174-6.

Katz, A. (2008). Intelligent agents and internet commerce in ancient Rome. *Society for computers and law, 20,* 35–38.

Kopelman, L. M. (2007). Using the best interests standard to decide whether to test children for untreatable, late-onset genetic diseases. *The Journal of medicine and philosophy, 32*(4), 375–394.

Kuehnle, W. H. (1988). Secondary Liability under the Federal Securities Laws - Aiding and Abetting, Conspiracy, Controlling Person, and Agency: Common-Law Principles and the Statutory Scheme. *Journal of Corporation Law, 14*(2), 313–376. Retrieved from https://heinonline.org/HOL/P?h=hein.journals/jcorl14&i=323. Accessed 12 July 2020.

MacCormack, G. (1974). INHERITANCE AND WERGILD IN EARLY GERMANIC LAW: II. *Irish Jurist (1966-), 9*(1), 166–183. Retrieved from www.jstor.org/stable/44026303. Accessed 12 July 2020.

Mason, E. (2018). Value Pluralism. *The Stanford Encyclopedia of Philosophy.* Spring 2018. Retrieved from https://plato.stanford.edu/archives/spr2018/entries/value-pluralism/. Accessed 8 Jan 2021.

Mill, J. S. (1859/2015). *On Liberty, Utilitarianism, and other essays*. Oxford University Press.

Miller, J. D., Yampolskiy, R., & Häggström, O. (2020). *An AGI Modifying Its Utility Function in Violation of the Orthogonality Thesis*. Retrieved from https://www.researchgate.net/publication/339642009. Accessed 12 July 2020.

Moore, R. H. (1988). Civil Liability for Negligent and Inadequate Training: A Private Security Problem. *Journal of Contemporary Criminal Justice, 4*(2), 106–118. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/104398628800400205.

Musiał, M. (2017). Designing (artificial) people to serve–the other side of the coin. *Journal of Experimental & Theoretical Artificial Intelligence, 29*(5), 1087–1097.
Nyholm, S. (2020). *Humans and robots : ethics, agency, and anthropomorphism*. Rowman & Littlefield Publishing Group.
Omohundro, S. M. (2008). The Basic AI Drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial General Intelligence, 2008: Proceedings of the First AGI Conference* (pp. 483–492): IOS Press.
Pagallo, U. (2010). Robotrust and legal responsibility. *Knowledge, Technology & Policy, 23*(3–4), 367–379.
Petersen, S. (2011). Designing People to Serve. In P. Lin, G. Bekey, & K. Abney (Eds.), *Robot Ethics* (pp. 283–298). MIT Press.
Petersen, S. (2017). Is It Good for Them Too? Ethical Concern for the Sexbots. In J. Danaher & N. McArthur (Eds.), *Robot Sex: Social and Ethical Implications*.MIT Press.
Phillpotts, B. S. (1974). *Kindred and clan in the Middle Ages and after; a study in the sociology of the Teutonic races*. Octagon Books.
Prinz, J. (2003). Level-headed mysterianism and artificial experience. *Journal of Consciousness Studies, 10*(4–5), 111–132.
Rini, R. (2017). Raising good robots. *aeon*. Retrieved from https://aeon.co/essays/creating-robots-capable-of-moral-reasoning-is-like-parenting. Accessed 12 July 2020.
Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI, 5*(15). Retrieved from https://www.frontiersin.org/article/10.3389/frobt.2018.00015.
Schaerer, E., Kelley, R., & Nicolescu, M. (2009). *Robots as animals: A framework for liability and responsibility in human-robot interactions.* Paper presented at the RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication.
Schulzke, M. (2013). Autonomous Weapons and Distributed Responsibility. *Philosophy & Technology, 26*(2), 203–219. Retrieved from doi:https://doi.org/10.1007/s13347-012-0089-0
Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. *Midwest studies in philosophy, 39*(1), 98–119.
C.F.R. Title 17. Commodity and Securities Exchanges. (1987).
Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy, 24*(1), 62–77. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-5930.2007.00346.x.
W Totschnig 2020 Fully Autonomous AI Retrieved fromdoi:https://doi.org/10.1007/s11948-020-00243-z
Turner, J. (2019). *Robot rules*. Palgrave Macmillan.
Walker, M. (2006). A moral paradox in the creation of artificial intelligence: Mary Poppins 3000s of the world unite! In T. Metzler (Ed.), *Human implications of human-robot Interaction: Papers from the AAAI workshop* (pp. 23–28). AAAI Press.
White, T. N., & Baum, S. D. (2017). Liability for Present and Future Robotics Technology. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot Ethics 2.0* (pp. 66–79). New York: Oxford University Press.
Wigmore, J. H. (1894). Responsibility for Tortious Acts: Its History. *Harvard Law Review*, 315–337.
Young, C. W. (1990). Respondeat Superior: A Clarification and Broadening of the Current Scope of Employment Test. *Santa Clara Law Review, 30*, 599.