Robots, Law and the Retribution Gap

By John Danaher

Forthcoming in *Ethics and Information Technology*

Abstract

We are living through an era of increased robotisation. Some authors have already begun to explore the impact of this robotisation on legal rules and practice. In doing so, many highlight potential liability gaps that might arise through robot misbehaviour. Although these gaps are interesting and socially significant, they do not exhaust the possible gaps that might be created by increased robotisation. In this article, I make the case for one of those alternative gaps: the retribution gap. This gap arises from a mismatch between the human desire for retribution and the absence of appropriate subjects of retributive blame. I argue for the potential existence of this gap in an era of increased robotisation; suggest that it is much harder to plug this gap than it is to plug those thus far explored in the literature; and then highlight three important social implications of this gap.

Keywords: Robotics; Law; Moral Responsibility; Liability Gaps; Retribution Gaps

1. Introduction

We are living through an era of increasing robotisation. Robots are fighting our wars, manufacturing our goods, stacking our warehouses, and caring for our most vulnerable citizens. Soon they will be driving our cars, delivering our goods, cooking our meals, and generally taking over large swathes of human activity (Ford 2015; Kaplan 2015). Technological changes of this sort have profound social, moral and legal implications. This article takes a narrow look at the potential impact of advanced roboticisation on our attitudes toward punishment and wrongdoing.

This is not a new area of inquiry. Legal theorists and philosophers have long thought about the potential impacts of robots on law and morality. Many have identified gaps in the existing legal-regulatory infrastructure that are challenged by the rise of the robots (Calo, Kerr and Froomkin 2016). But most of these identified gaps focus on liability issues that arise from robot misdeeds (i.e. who should be liable if a robot injures or harms another human being?) (Matthias 2004; Calo 2015), or more fancifully on the philosophical question of whether a robot could be morally and legally responsible (Matthias 2004; Purves et al 2015). What these pre-existing inquiries have missed is the potential 'retribution gap' that could arise from the widespread use of robots. In this article, I try to clarify and make the case for this gap.

My argument is simple. Psychological evidence suggests that humans are innate retributivists (Carlsmith and Darley 2008; Jenson 2010): when they are harmed or injured they look for a culpable wrongdoer who is deserving of punishment. Many legal and moral philosophers argue that this retributive attitude is the correct one to take (Alexander and Ferzan 2009; Moore 1993; Duff 2007). Increasing levels of robotisation make it likely that robots will be responsible for more and more harm and injury, but the robots themselves are unlikely to meet the conditions for retributive blame. Consequently a retribution gap is opened up: people will be eager to find an appropriate and deserving subject of retributive punishment, but none will be found. This gap could have a number of significant social and legal implications.

The article defends this argument in four parts. First, I clarify the conceptual terrain and distinguish more clearly between 'liability gaps' and 'retribution gaps'. Second, I introduce and defend the argument for thinking that increased robotisation

will give rise to a retribution gap. Third, I reply to objections to this argument. Fourth, and finally, I consider three important social implications of the retribution gap.

2. Defining Robots and the Retribution Gap

The argument in this article has to do with the nature of *robots,* the phenomenon of *robotisation* (i.e. the increasing use of robots), and the impact of both on social practices of responsibility and blame. It is important to start by clarifying the conceptual terrain associated with these phenomena. Doing so allows me to situate the argument I defend relative to somewhat similar arguments.

I start by clarifying that my argument is concerned with robots that have a high degree of autonomy. What it means for a robot to be autonomous is a matter of some debate in the literature on robot ethics (e.g. Sparrow 2007, 64-66). Here I adopt the approach of Hellstrom (2013). He argues that we can distinguish between robots based on their 'autonomous power'. He defines this as a gradient-concept which denotes "the amount and level of actions, interactions and decisions an agent is capable of performing on its own" (Hellstrom 2013, 101). In other words, it denotes the ability of the robot to act in the world without input or control from a human designer, programmer or operator. Hellstrom intends for the concept to have a broad potential application, but for degrees of autonomous power to be discernible. The more actions and the more diverse the range of environments in which they can be deployed without human interference or control, the more autonomous the robot is. So, for example, Hellstrom argues that a landmine has a very low degree of autonomous power: it is capable of performing one 'action' (detonation) in a range of environments, in response to pressure/mass on its triggering mechanism, without the need for human control. A self-driving car would have a much higher degree of autonomous power: it would be able to act in various ways (breaking, turning, accelerating etc), across a range of environments, without the need for human interference or control. Contrariwise, certain objects that we often deem to be robotic can have no autonomous power whatsoever. Teleoperated military drones are like this because they rely on human input and oversight to exert their causal powers.

The stipulation that robots have a high degree of autonomous power is relevant to my argument because it is that power that threatens to break the link between the human

creators and designers of such systems and their ultimate causal effects. If robots were little more than tools — as teleoperated systems effectively are — then there would be no risk of a 'liability' or 'retribution' gap opening up in the law. The human controllers would be liable and subject to retributive blame. The fact that robots are being created with relatively high degrees of autonomous power does threaten to open up such gaps.

This brings us to the second key conceptual clarification: the distinction between a liability-gap and a retribution-gap. This is a distinction that has been neglected in the literature to date and is central to the argument I present in this article. In distinguishing between these two concepts it is worth keeping in mind a set of background concepts. These concepts all centre around the notion of moral and legal responsibility. 'Responsibility' is a bundle-concept and the term can be used in a number of distinct ways. In saying this, I follow the lead of HLA Hart (1968) and, more recently, Nicole Vincent (2011). 'Responsibility' is a term used to denote the relationship between an agent, its actions, and the outcomes of those actions. There are several such relationships that are relevant from a legal and moral perspective.[1] Three of those relationships are important to the present argument. The first is that of *causal responsibility*, which denotes a causal link between the agent, their actions, and some particular outcome. The second is that of *moral/legal responsibility*, which denotes the fact that the causal link between the agent and the action/outcome is such that the agent is an appropriate subject of legal/moral blame.[2] This is usually determined by whether the agent has the right *capacities* and whether those capacities were exercised at the relevant time. The third is *liability responsibility*, which denotes the punishments or sanctions that an agent must bear in virtue of its *moral/legal* responsibility. The concept of liability responsibility can be further distinguished depending on the relevant area of moral/legal practice. Thus, in legal practice, distinctions can be drawn between compensatory-liability (which applies primarily in civil/tort law and sometimes in criminal law) and is about paying back the victims of harm or injury; and punitive-liability (which applies primarily in criminal law) and is about suffering harm and public condemnation for wrongs done. This means that there are several sub-types of liability gap and the retribution gap is one such sub-type – one that is not associated with compensation or burden-sharing.

[1] Vincent (2011) argues that there are at least six: virtue-responsibility; role-responsibility; outcome-responsibility; causal-responsibility; capacity-responsibility; and liability-responsibility.
[2] The concepts of legal and moral responsibility can be distinguished and sometimes pull apart. However that possibility can be ignored here.

In the case of human agents, causal-, moral/legal- and liability- responsibility tend to come together. If I drive a car whilst drunk, and if in doing this I collide with and injure a pedestrian, then I am *causally* responsible for the injury, I am legally and morally blameworthy for causing the injury, and I am forced to pay compensation and serve time in jail. The problem with the rise of robotisation is that these links can break down. A robotic agent, with the right degree of autonomous power, will tend to be *causally responsible* for certain injurious or harmful actions. However, the robot will not be morally and legally responsible (because it will lack the requisite moral capacities), nor will the human creators and designers be morally/legally responsible because the robot has a sufficient level of independence from them. The end result is a liability gap: there is no suitable agent who can bear the burden associated with the injurious outcome. But the precise nature of the gap can vary depending on particular legal practice. Thus, in the civil law context the gap will tend to be compensatory in nature, whereas in the criminal context it will be retributive in nature.

This might seem a little sketchy. How exactly do the gaps arise? The answer is that the gaps arise when the moral/legal tests for determining who should bear the burden fail to align with the reality of who is causing the injurious outcome. This argument has been specifically traced out by several writers in relation to the compensatory gap that could arise from the use of robots. Calo's recent discussion is instructive (2015, 129-131). As Calo points out, civil law tests for liability typically require a plaintiff (i.e. victim of injury) to prove that (i) the defendant owed them a *duty of care* and (ii) that the defendant breached a *standard of care*. It may be relatively easy to argue that a robot manufacturer or designer owes a customer or third party a duty of care. Indeed, some of the most famous cases in legal history make the existence of such a duty clear.[3] The difficulty arises with the standard of care. Many legal tests insist that the injury suffered by the plaintiff be reasonably foreseeable by the defendant. The problem is that robots are increasingly programmed with machine learning algorithms that lead them engage in acts that are not anticipated, expected or reasonably foreseeable by the original manufacturer. The traditional legal standard cannot be stretched to cover the kinds of scenario made possible by advanced autonomous robots.

[3] *Donoghue v Stevenson* [1932] AC 562 - Is a foundational decision in English tort law holding that you owe a duty of care to your 'neighbour', where neighbour is defined relatively broadly. In that particular case, it included the consumer of a product who was not its actual purchaser.

The result is a liability gap: there is an injurious outcome but no legally identifiable compensation-giver.

Compensation gaps of this sort are interesting and worthy of consideration. Nevertheless, there are alternative civil legal standards that can be used to plug those gaps. For example, a combination of vicarious liability rules and strict liability rules could do the trick.[4] These would allow one person to be responsible for the actions of another and do away with the need to prove reasonable foreseeability. Indeed, Calo himself suggests that increased robotisation could give rise to increased use of strict liability standards (2015, 40; Scherer 2016 makes a similar point about AI). Alternatively, there could be greater use of social insurance funds to pay out compensation to victims of robotic harm. So although this is something worth worrying about, there are plausible ways to solve the problem and ensure that people receive the necessary compensation.

What I want to argue in the remainder of this article is that increased robotisation can also give rise to important gaps in criminal liability, specifically gaps associated with the attribution of retributive blame for wrongdoing. Retributivism is a theory of punishment which holds that people should be punished (i.e. suffer some harm or setback to interests) for wrongdoing because they *deserve* to be punished. Embedded in this is the notion that their *desert* is a function of their moral culpability for their actions (Moore 1993). The retributive gap arises when people look for culpable wrongdoers to blame for some injurious outcome but none can be found. This gap has been neglected in the existing literature on robots and the law, and is much less easy to plug.

It is important to justify this claim by separating the argument I am about to present from other similar arguments. Matthias (2004) for instance argues at length in favour 'responsibility' gaps that arise from the growth of machine learning automata. And he makes a strong case for thinking that such automata will not be appropriate bearers of responsibility, but in doing so he doesn't distinguish between the differential significance of liability responsibility in different legal contexts. His comments suggest a concern with who will pay for the wrongdoing of robots, not so much with who deserves retributive blame. Highlighting the importance of this distinction is part of the

[4] There are also liability standards associated with control and care for animals that might be adopted by analogy.

present goal. Sparrow (2007) comes closer to the argument I wish to defend by looking at responsibility gaps in the deployment of autonomous weapons systems. He argues that the principles of just war require that there be a morally responsible agent making lethal decisions about legitimate targets, and that highly autonomous weapons systems undermine this requirement. Sparrow's argument has given rise to a rich debate about whether the use of autonomous weapons systems really does breach the conditions of just war (e.g. Purves et al 2015; Simpson and Muller 2016), but the argument differs from the one I present in three important respects. First, I am not concerned with military cases nor with the principles of just war; I am concerned with more mundane and everyday uses of autonomous robots. Second, the argument I make is not primarily an ethical one. I am not objecting to the use of autonomous robots, nor do I think their development is a bad thing. Instead, I am arguing for a mismatch between certain psychological desires for punishment and normative theories of punishment. Third, my goal is to consider the broader social and legal consequences of this mismatch, not to claim (as both Matthias and Sparrow do) that this is a gap that urgently needs to be filled. That said, the argument I defend uses some similar concepts and ideas – I will identify these in the following section and further highlight the differences.

3. The Argument for the Retribution Gap

My argument for the retribution gap works like this:

(1) If an agent is causally responsible for a morally harmful outcome, people will look to attach retributive blame to that agent (or to some other agent who is deemed to have responsibility for that agent) — what's more: many moral and legal philosophers believe that this is the right thing to do.

(2) Increased robotisation means that robot agents are likely to be causally responsible for more and more morally harmful outcomes.

(3) Therefore, increased robotisation means that people will look to attach retributive blame to robots (or other associated agents who are thought to have responsibility for those robots, e.g. manufacturers/programmers) for causing those morally harmful outcomes.

(4) But neither the robots nor the associated agents (manufacturers/programmers) will be appropriate subjects of retributive blame for those outcomes.

(5) If there are no appropriate subjects of retributive blame, and yet people are looking to find such subjects, then there will be a retribution gap.

(6) Therefore, increased roboticisation will give rise to a retribution gap.

This argument is structurally straightforward. It chains together two sub-arguments. The first sub-argument is about the desire for retributive blame and how it may look toward robotic agents; the second sub-argument is about how that desire will go unfulfilled. In what follows, I offer an initial clarification and defence of the premises. I consider a range of objections along the way, but defer two objections to the next section of the article.

Let's start by looking at premise (1). The crucial concept in this premise is that of retributive blame. Here, I adopt a standard account of retribution and retributive blame (Moore 1993; Boonin 2008; Zimmerman 2011; Kramer 2011). I view retributivism as the belief that agents should be punished, in proportion to their level of wrongdoing, because they *deserve* to be punished. I view retributive blame as being appropriate when the agent is morally culpable for the harm that occurred. Culpability is a function of a number of standard legal and moral tests. Most typically, an agent is culpable for criminal wrongdoing if they deliberately intended some moral harm; or were recklessly indifferent or grossly negligent with respect to that moral harm (Moore 1997). I take it that lesser standards of culpability (e.g. mere negligence as opposed to gross negligence) would not be suitable for retributive blame. I accept that one agent could attract retributive blame for the actions of another agent if they have sufficient control and/or influence over that agent's choices.

This should be relatively uncontroversial. The more problematic aspect of premise (1) concerns its awkward dance between descriptivity and normativity. This awkward dance is critical to the argument and will continue for the remainder of the article. On the one hand, the premise appeals to the notion that there are powerful psychological drives pushing people to locate appropriate subjects of retributive blame. This is a descriptive/predictive claim. On the other hand, the premise appeals to the notion that

some moral and legal philosophers think this is the right thing to do. This is not a normative claim *per se* but it is an implicit appeal to the fact that many theorists think the retributive attitude has strong normative grounding and so should provide the underpinning for our criminal justice system.[5] The combination of this normative stance with the general social desire for retribution is what makes the retribution gap worthy of our attention (or so I shall argue). As I will argue, it is the potential mismatch between the general desire for retribution and the specific requirements of retributive moral theory that makes the retribution gap particularly disturbing. To make this point I need to first defend the claim that there is a general social desire to engage in retributive punishment.

Three pieces of evidence support that claim. The first is the human tendency to attribute events to acts of agency, even when they are not actually acts of agency. Humans appear to have hyperactive agency detection devices (HADDs) in their brains (Barrett 2004; Boyer 2002; and Atran 2002). Barrett (2004) argues that the HADD kicks-in whenever non-inertial movement is detected in our surrounding environments. This has sometimes been cited as a major explanation for religious beliefs (particularly traditional pagan/animist beliefs), and there are a number of explanations for why humans would have a tendency to interpret events as the products of agency. This tendency is not essential to the argument I am making, but it is supportive insofar as it suggests that humans will be inclined to view the harmful outcomes of robotic action as being a product of agency. This tendency opens the door to the attribution of punishment and blame.

The second bit of evidence has to do with the human proclivity to punish. Ethnographic evidence suggests that practices of punishment and blame are common to all human societies (Jensen 2010; Brown 1991). What's more, experiments reveal that humans have a strong tendency to punish anyone they believe to be violating group norms. Indeed, experiments reveal that they will do this even when it is costly to themselves (Gintis 2011, Ch 3.6). This proclivity for punishment is further underscored by neurobiological evidence suggesting that punishment activates parts of the brain's reward circuit and so is likely to feel pleasurable (Jensen 2010; Pinker 2011, 529-532).

[5] There are many criminal theorists who support this basic position: Michael Moore (1993; 1997); Larry Alexander and Kimberly Ferzan (2006); and Antony Duff (2007). These theorists support the view on moral/philosophical grounds and could be classified as pure retributivists; others support it in part because it is the dominant social/psychological attitude, e.g. Robinson and Kurzban (2007).

There are a variety of explanations for the tendency to punish, but they are not relevant here. All that matters is that there is this tendency to punish when social norms have been violated.

The third piece of evidence is the fact that when people punish they tend to do so in accordance with the criteria of retributivism and not the criteria of alternative theories of punishment like deterrence or rehabilitation. Support for this comes from experimental work and from the evolutionary and ethnographic record (Carlsmith and Darley 2008; Jensen 2010). The experimental work is particularly instructive. Through a series of studies, Carlsmith and Darley (2008) have revealed six important lines of evidence supporting the claim that people are 'natural' or 'innate' retributivists. That is to say, they are inclined to punish people in a manner that is (a) *proportionate* to their level of wrongdoing and (b) *sensitive* to their degree of *blameworthiness*. The six lines of evidence include the fact that people seem to be more sensitive to retributive criteria, more attracted to bits of evidence that are relevant to retributive modes of punishment, better able to understand retributive theories of punishment, and are unlikely to support an alternative system of punishment (e.g. restorative justice) that does not include retributive criteria. In addition to this, it is found that there is often a gap between self-reported attitudes to punishment and actual behaviour (i.e. people might claim to favour deterrence but in practice favour retribution) and even within diverse experimental populations it is rare to find people with a consistently non-retributive approach to punishment. Collectively, these lines of evidence provide strong support for the claim that people are inclined to punish in accordance with retributive criteria.

Let's turn attention to premise (2). This premise claims that increasing robotisation will lead to increasing levels of harm being caused by robot agents. This seems like an obvious truth: if robots participate in more and more activities, and if they have more and more autonomous power, then it is likely that they will (at least on occasion) be causally responsible for moral harm. A self-driving car, or an autopilot mechanism, or even a robot waiter that miscalculates at an inopportune moment could cause injury or death. Indeed, there are examples of this already happening. In July 2015, the *Financial Times* reported that a German worker had been killed by a robot in a car manufacturing plant (Bryant 2015). Some people referred to this as the first act of robot homicide. Whatever the merits of that attribution, incidents of this sort can be expected to multiply in line with the increase in robotisation.

Someone might object to this line of reasoning on the grounds that robots could be morally perfect and hence highly unlikely to be causally responsible for moral harm. Indeed, safety and reduced risk is often one of the major rationales behind robotisation. Google for instance have explicitly argued that their self-driving car should come without a steering wheel because it is when humans interfere with the robot that accidents are most likely to occur (Walker 2016). But this objection seems naive for at least three reasons. First, even if robots are less likely than humans cause harm they are unlikely to be perfect. Even if the probability of robot-caused harm is minimal this would still translate into an increased amount of causal responsibility for moral harm if robots participate in more and more activities. Second, it is naive to assume that the creation of even morally excellent robots (let alone morally perfect robots) is easy or straightforward. It is very difficult to program or train a robot to follow the kinds of moral rules we would like. The difficulty of this task is one of the things that has spurred the recent debate about AI risk, and the doomsaying pronouncements of tech gurus like Elon Musk and Bill Gates (Sainato 2015). There are several reasons why it is so difficult to get robots to engage in appropriate moral behaviour. One is that we don't agree ourselves on what appropriate moral behaviour is in all contexts; another is that most moral rules admit of counterexamples or exceptions, particularly if followed literally or in unexpected ways, as may be likely in the case of robots — it is probably impossible to foresee and avoid all those exceptions; and another is that trying to train robots to learn moral rules, through some machine learning algorithm, will often generate unexpected results as the robot extrapolates a rule from an unappreciated feature of the environment (Muehlhauser and Helm 2013). These and other problems are widely discussed in the literature and seem to add support for premise (2) (Matthias 2011). Third, there may be no incentive to create a morally perfect robot. In fact, some robots may be created in order to engage in morally circumspect behaviours, or could have powers and abilities that render their behaviour immoral in certain contexts.

If premises (1) and (2) are accepted, then premise (3) would follow and we reach the interim conclusion that in a world of increased robotisation people are likely to look to robots (or associated agents like manufacturers and programmers) as potential subjects of retributive blame. This brings us to premise (4). This one claims that neither robots nor associated agents like manufacturers or programmers will be appropriate subjects of retributive blame. It is important that this premise is properly interpreted.

When I say that these agents will not be 'appropriate' subjects of retributive blame, I mean that they are not appropriate in a normative sense. People may very well attach blame to these individuals — a possibility to which I return — but they will be *wrong* to do so. That said, there is a descriptive element to the claim too. I suspect that even if people do attach retributive blame to such agents, they will tend to be unsatisfied by the results. Furthermore, the claim is intended to be a relatively modest one. I am not suggesting that it will *never* be appropriate to attach blame to a robot or to its manufacturers/programmers. In some cases it may be right to do so. I am simply claiming that it will be difficult to do so as robots attain more autonomous power.

Premise (4) can be broken down into two parts. The first part claims that robots themselves will not be appropriate targets of retributive blame; the second part claims that robot manufacturers and programmers will not be appropriate targets of retributive blame. Similar premises have been defended in the literature before (Matthias 2004; Sparrow 2007). Here, I adapt these defences to the present argument.

I start with the robots themselves. Recall that retributive blame requires culpability. Culpability is a function of both causal and mental/moral responsibility for the outcome. In other words, the outcome must be physically brought about by the agent, the agent must have the right kind of mental capacities that open them up to blame, and they must have exercised those capacities at the relevant time (Vincent 2011). The mental capacities for being blamed are traditionally understood in terms of various intentional states (i.e. beliefs, desires, intentions). So, for example, the agent must *know* that their actions will (or could) bring about some morally harmful outcome, and they must either intend or be reckless or grossly negligent with respect to that outcome. Some even argue that conscious representation of the relevant beliefs, desires and intentions is necessary for blame (Levy 2014).

I assume here that robots can be causally responsible for certain outcomes. The tricky question is whether they can have and exercise the requisite mental capacities (Gunkel 2012; Asaro 2011). There are several reasons to doubt that they can. The first is that there are long-standing critiques of the notion that a programmed cognitive architecture can replicate or instantiate the kinds of conscious mental state that many deem necessary for responsibility (Purves et al 2015). Long-standing objections to the notion of 'Strong AI' hold that such created artifacts can never have the original

intentionality that is required for human-like mental processing. We can speak, analogically, of robots *desiring* certain outcomes but we should doubt whether they actually desire outcomes in anything like the way we do, or that they do so in the way we deem necessary for retributive blame. The second reason is that we already know that existing AI architectures work best when they do not replicate human-like mental architectures. Chess-playing computers and facial-recognition programs do not approach their tasks by following human-like strategies or methodologies. They work by exploring unfathomably large datasets and extrapolating rules and strategies from those datasets. In doing so, they can be instrumental reasoners par excellence, but, as Bostrom (2012; 2014) points out, there is no reason why their approach to such tasks would involve the functional analogues of mental states like beliefs and/or desires and intentions. Programs could be constructed so as to follow optimisation processes that lack sub-components that line up with what we call beliefs or intentions. And even if they could have human-like mental architectures, these may not include moral faculties or sensitivity to moral reasons for action. A final for reason for doubt has to do with past attempts to ascribe blame to non human-like agents (List and Pettit 2011). The best example of this comes from recent attempts to ascribe criminal liability and blame to corporations. Such attempts typically boil down to fines or dissolutions, and to additional punishments of the individuals who run these corporations. People are generally unsatisfied with penalties ascribed solely to the corporate agents. This was noticeable in the aftermath of the 2008 financial crisis when there were extensive public calls for individual bankers and CEOs to suffer punishment; not simply for their companies to be fined or dissolved.

The idea that we could ascribe retributive blame to the robots themselves is philosophically interesting — and I will return to it below — but given these problems I suspect it is far more likely that when a robot misbehaves people will look to the human manufacturers and programmers as potential targets for retributive blame. This is certainly the dominant assumption in the existing literature, and the approach recommended by some (Calo 2015; Hellstrom 2013; Chisan Hew 2014). In the case of simple rule-following robots, with a handful of creators, and a limited degree of autonomous power, this might be straightforward enough. But with anything more sophisticated, two problems will start to emerge: (i) the level of robotic autonomous power may be such as to break the link needed for vicarious blame attribution and (ii) even if there is some link, the degree of blame is likely to be seriously attenuated,

meaning that there is a level of harm that is unmatched by a proportionate or corresponding level of retributive blame. Either way, there is a 'gap' in the potential application of retributive blame.

As to the first problem, we already see ways in which machine learning can give rise to emergent behavioural patterns that are unanticipated and unexpected by the original programmers. Such acts of autonomous creativity are likely to increase as machine learning programs get better. And the net effect is likely to be compounded as robotic cognitive architectures are assembled from pre-existing packages of code and grafted onto complex algorithmic ecosystems (Kitchin 2016). The programmers and manufacturers will consequently neither intend nor be reckless with respect to the potential misbehaviours of their robotic creations: the robots will learn to think and act in ways that are beyond the intentions and expectations their original creators. Retributive blame for those creators will therefore be blocked.

This is in turn connects with the attenuation problem. Even if there is some residual link between the creators and the robot, it is likely to be attenuated to the point where it would not be morally appropriate to ascribe a level retributive blame to those creators that covers the full gravity of the moral harm done by the robots. This attenuation problem will be further compounded by the fact that sophisticated robots with autonomous power are likely to be created by large teams of programmers and designers, none of whom have individual control or responsibility for the final robot. At best then you have a distribution of an attenuated level of blame across a broad number of individuals.

Someone might respond at this point and claim that the full level of retributive blame could be ascribed to the manufacturers and programmers if we simply broaden our understanding of what it means to be reckless or grossly negligent with respect to robotic behaviours. Perhaps it is reckless or grossly negligent to create any machine, with a high degree of autonomous power, that could be causally responsible for moral harm? Perhaps we can always ascribe retributive blame to the creators of such robots. But this looks like an unwelcome suggestion. First, note that we don't do this for the misuses of other created devices, particularly if the devices have potentially positive uses. Second, it would probably unfair to do so if the behaviours of the robots are truly beyond the reasonable expectations of their designers. Fairness and proportionality are

key aspects of the retributive philosophy: you give people what they deserve, nothing more or less.

That brings us to premise (5). This one doesn't need to be defended so much as properly explained. The claim is that when you combine a general desire to find appropriate targets of retributive blame, with the fact that no such targets can be found, you get a *retribution gap*. This is where the awkward dance between descriptivity and normativity reemerges. The gap being mooted is a gap between what is desired and what some people believe to be, morally speaking, right. As mentioned earlier, it is the mismatch between the normative and descriptive that I find most interesting because it is this mismatch that gives rise to the more significant social and legal consequences of the retribution gap. But a moral retributivist could dispute premise (5)'s claim about the existence of a normatively important gap. Retributivism is the belief that people should be punished because they deserve it. Thus, for a retributivist, a retribution gap can only arise when there is an appropriate target of retributive blame to whom blame is not being ascribed. But what I am claiming in this argument is that in the case of robot-caused harm, there may be no appropriate targets of blame. But in that case there's no 'gap' that should concern the moral retributivist. Nobody who deserves punishment is going unpunished.

This view is correct insofar as it accurately states what is normatively relevant from the perspective of a retributivist. But that does not mean that there are no moral/normative problems arising from the gap between what is desired and what is retributively appropriate. As I shall point out in the final section, the gap between what is desired and what is appropriate gives rise to a number of normative concerns, including concerns about moral scapegoating, that should be of interest to everyone, even the most staunch of retributivists.

With this explanation of premise (5) out of the way, the initial defence of the argument is complete. If the argument is right, then the increase in robotisation will lead to an increase in the causal responsibility of robots for morally harmful outcomes. Since humans are naturally inclined to find someone to retributively punish when morally harmful outcomes occur, this will lead to people desiring some appropriate target of retributive blame for acts of robot harm. But since, in many cases, neither the robots nor

the manufacturers/programmers will be appropriate targets of retributive blame, a 'gap' will open up. A desire for retribution will go unfulfilled.

4. Objections and Replies

It is worth singling out two further objections and subjecting them to closer scrutiny. The first is the *Anthropomorphisation Problem* and the second is the *Command Responsibility Objection*. Both cast doubt on the likely existence of a retribution gap, or suggest easy ways in which it can be plugged.

The first objection can be illustrated by reference to a particular scene in the popular BBC sitcom *Fawlty Towers*. The series focuses on the eccentricities and misfortunes of Basil Fawlty, the owner and manager of a small hotel in Torquay, England. In the episode 'Gourmet Night', Basil tries to attract the local upper class to his hotel by hosting a gourmet dining experience. Unfortunately for him, his head chef gets drunk, and he has to source the food for the evening at another restaurant. On the drive to the other restaurant, his car breaks down and, clearly at the end of his tether, he proceeds to shout at it and to give it a 'damn good thrashing' with a fallen-down tree branch. In this sense, he appears to blame the car for his misfortune and to mete out some punishment to it. The scene illustrates an attitude that humans might take when robots misbehave. They might react like Basil Fawlty did to his car. They might anthropomorphise the robot — i.e. falsely ascribe relevant human moral faculties to the robot — and then feel comfortable punishing it in much the same way that Basil Fawlty felt comfortable thrashing his car. In this manner, the alleged retribution gap would vanish: the desire for retribution would be satisfied through the process of anthropomorphisation.

There is no doubt that we tend to anthropomorphise technological artifacts and that when we do so we start to behave towards them as we would another human being. Indeed, this phenomenon has already been researched by those interested in human-robot interactions. Several studies suggest that humans are willing to ascribe responsibility to robots in certain contexts. Some specific findings from these studies are particularly interesting. The first, from a study by Kim and Hinds (2006), found that humans would ascribe responsibility to a robot delivery machine but that this depended on the degree of autonomous power the robot had, i.e. the more power, the more likely

the ascription of responsibility. This finding might suggest a further problem for my argument insofar as it could be the case that at low levels of autonomous power the possibility of vicarious blame-attribution (i.e. blaming of manufacturers) is acceptable; while at higher levels the robots themselves become the targets; but at no stage is there a 'gap' in the human willingness to assign blame. But this has to be tempered by the fact that studies also find that willingness to assign blame is dependent on other seemingly less relevant factors. For instance, Kim and Hinds (2006) found that blame-attribution was lessened if the robot was more transparent about what it was doing — i.e. if it explained to the humans what it was trying to do. And in another study Hinds et al (2004) found that robots were deemed less responsible the less humanoid they were. These two factors should, arguably, be irrelevant to blame-attribution: what should matter is whether the robot had the requisite capacities and whether it exercised them at the relevant time, not whether it looked like a human or whether it told people what it was going to do. Other studies have found similar effects (Marin, A et al 2013; Kiesler & Goetz 2002)

There are several reasons to doubt whether the tendency to anthropomorphise can cover the alleged retribution gap. If all we care about is whether the desire for retribution is fulfilled, then by all means we can take advantage of this tendency to anthropomorphise. We can study the quirks and biases of human blame-attribution, and design robots that cater to these quirks and biases (e.g. make the robots less transparent, but more humanoid). But if we take this approach, we should be aware of potential impediments. Not every robot manufacturer will have the incentive to create a humanoid robot. The incentive might exist when creating robot carers — because the manufacturers want the human users to feel comfortable with the carers[6] — but whether it exists in other industries is doubtful. There is no real need for a robotic car or military drone to take on a human-like form. And yet those kinds of robots might be ones that are causally responsible for the most harm. So unless we actively force the manufacturers of such devices to make robots that cater to human blame-attribution quirks, desires for retribution may remain unfulfilled. But more importantly than this, even if we did implement such rules, a true retributivist should remain unhappy. In effect, all we are doing is tricking ourselves into believing that we have found

[6] Though note the potential impact of the so-called 'Uncanny Valley' effect - if the robots are too humanoid they may be too creepy for the human users. The uncanny valley was first hypothesised by Masahiro Moti in the 1970s and has recently been confirmed in some experimental tests, but how deep and wide the valley actually is remains contentious. See: MacDorman & Ishiguro (2006); MacDorman (2006); MacDorman, Green, Ho & Koch (2009).

appropriate targets of retributive blame. In fact, there is a great danger in going down this route. The one incentive that companies might have to create robots that cater to the quirks and biases of human blame-attribution processes would be that doing so could allow *them* to avoid being targets of blame attribution. This should be truly worrying if the manufacturers could be legitimate targets of blame.

This brings us to the *Command Responsibility Objection*. This one claims that the alleged retribution gap could be plugged if we simply change our attitude toward the manufacture and production of robots with high levels of autonomous power (Hellstrom 2013). We should view the process as being akin to that which takes place in the military. In the military, troops are responsible for implementing and carrying out the orders of their commanders. The commanders then take responsibility for any misdeeds by their troops in carrying out those deeds. We could look at the creation, manufacture and eventual 'release' of an autonomous robot in a similar fashion. If such devices are created and released by large companies, like Google, then the senior management within that company should have command responsibility for what happens when the devices are released. They can then be the appropriate targets of retributive blame. The gap can be plugged.

It is important to realise how this objection differs from those previously considered whilst defending premise (4) of the main argument. It is not claiming that we simply 'stretch' or adapt existing standards and tests of blame-attribution. It is arguing that we adopt a new standard. In this, it is similar to the suggestion by Calo that we make more use of strict liability standards when dealing with potential compensatory gaps. The idea is that we set up a new regime of responsibility norms that apply to any company or organisation that develops autonomous robots. Anyone who gets into that business will know that they have command responsibility for the actions of their robots. This has two potential benefits. On the one hand, it should encourage them to be more cautious about releasing potentially dangerous robots or to build in safety protocols (e.g. kill switches) before doing so; and on the other hand, if the norms of responsibility are clearly announced in advance, it makes them more appropriate targets of retributive blame should something go wrong.

This may work to plug the retribution gap, but we should be aware of three potential pitfalls. The first is simply that command responsibility doctrines can

sometimes fail to comply with intuitions of retributive justice. Indeed, one of the most famous cases in the history of the doctrine — the *Yamashita* (1946) trial — strikes many people as failing to do this (Prevost 1992). The case involved a Japanese WWII military commander, Tomoyuki Yamashita, being prosecuted and executed for war crimes committed by his troops in the Philippines. The verdict was controversial because Yamashita was deemed to have command responsibility for his troops despite the fact that there was a breakdown in communications and he was (allegedly) unaware of what happened. This has led to more relaxed doctrines being pronounced in subsequent years. There is a danger of something similar happening to robot manufacturers if a similarly strict standard of responsibility is applied to them: they may be legally punished but this may fail to align with what is retributively appropriate. Or, if a more relaxed doctrine is applied, there may once again fail to be a target of retributive blame. A second problem with this approach is that it may have a stultifying effect on the growth and development of robotics. It is important to bear in mind that developments in robotics can be socially beneficial. A self-driving car with a lower risk of accidents could reduce the number of deaths on our roads. But if we impose too high a standard of responsibility on the manufacturers of such devices, we may slow (or completely block) their development. We need to consider whether the existence of a retribution gap is sufficiently serious to warrant that risk of stultification. Finally, it is worth bearing in mind that information technology now enables people to develop robotics or AI projects with a limited organisational infrastructure. Robots may not be developed by the large, well-integrated commercial enterprises of the 20th century; they may be developed by culturally and geographically distributed networks, with no clear hierarchy or visible infrastructure (Scherer 2016). Whereas it is relatively easy to impose something like a command responsibility framework onto a well-integrated, large organisation; it is much more difficult to do so with fragmented and distributed organisations.

5. Why the Retribution Gap Matters

Suppose then that there is a retribution gap. Who cares? Does the fact that people look for appropriate targets of retributive blame, but none can be found, have any important social or legal repercussions? I close by highlighting three potentially important implications. The strength and significance of these implications varies

depending on your preferred theory of punishment or your overarching theory of social justice/morality.

The first implication is that the existence of a retribution gap can give rise to an increased risk of *moral scapegoating*. If there is a deep human desire to find appropriate targets of retributive blame, but none really exist, then there is a danger that people will try to fulfill that desire in inappropriate ways. Or, perhaps even more serious than this, that other social actors will take advantage of the desire in inappropriate ways. I have hinted at this risk several times in this article. I have noted how robot manufacturers could toy with the quirks and biases of human blame-attribution in order to misapply blame to the robots themselves, and I have noted how doctrines of command responsibility or gross negligence could be unfairly stretched so as to inappropriately blame the manufacturers and programmers. Anyone who cares about the strict requirements of retributive justice, or indeed justice more generally, should be concerned about the risk of moral scapegoating.

The second implication is that the existence of a retribution gap could pose a *threat to the rule of law*. According to some legal theorists (Robinson 2013; Robinson & Kurzban 2007; Robinson, Kurzban and Jones 2007) the majority of people have a reasonably fixed set of intuitions about what kinds of behaviours or outcomes are morally harmful and about how people should be punished for engaging in or causing these outcomes (in this they include preference for retribution). They argue that the rule of law can be undermined if legal systems fail to align with these intuitive judgments. If the legal system seems to be out of touch with what ordinary folk think is right, these ordinary folk will lose trust in the legal system and may resort to vigilantism in an effort to seek justice. The existence of a retribution gap could exacerbate this phenomenon. If people feel that *someone* deserves retributive blame for the harmful acts of robots, but our legal and moral systems are incapable of finding anyone, you will have a situation in which intuitive judgments are out of line with legal practice. This could begin to erode respect for the rule of law. Moral retributivists could respond here by saying that ordinary folk simply need to recalibrate their intuitive judgments and understand why no appropriate subject of retributive blame can be found. That is all well and good, but this still requires that we get to grips with this potential threat.

The third implication follows on from the second. The view defended by Robinson and Kurzban has been labelled 'Punishment Naturalism' by its critics (Brahman, Kahan and Hoffman 2010). To them, Robinson and Kurzban's view assumes too readily that judgments of wrongdoing are fixed by a common and innate set of intuitions. Although there may be some consistency in such judgments in particular communities or states, this consistency is not natural or fixed. On the contrary, it is culturally contingent and open to being changed. Instead of Punishment Naturalism, these critics adopt a theory they call Punishment Realism, which openly acknowledges and respects the contingency and fluidity of our intuitions about wrongdoing and punishment. This is significant in the present context because one thing this criticism helps to highlight is how the existence of a retribution gap presents a *strategic opening for those who oppose retributivism*. An increased amount of robot-caused harm, in the absence of retributive blame, could shock or unsettle the cultural status quo. Since that status quo seems to be dominated by retributivism (in many countries), something needs to be inserted into the gap in order to restore the equilibrium. Those who prefer and advocate for non-retributive approaches to crime and punishment could find themselves faced with a great opportunity. Their calls for a more consequentialist, harm-reductionist approach to our practices of punishment and blame could have a better hearing in light of the retribution gap. Consequently, there is something of significance in the retribution gap for those who completely reject the retributivist philosophy.

6. Conclusion

In this article I have made three arguments. First, I have argued that debates about robotisation in the law need to look beyond its potential impact on doctrines of (civil) liability. In particular, they need to look beyond what I call 'compensation gaps' and how to plug them. Although these gaps are undoubtedly interesting and significant, they are relatively easy to plug. Second, I have argued that increasing robotisation could give rise to a far more interesting gap when looked at from the perspective of criminal liability. In particular, I have argued that more and more robots, engaging in more and more potentially harmful activities, could give rise to a 'retribution gap'. When people are harmed by the activities of a robot, they will look for potential targets of retributive blame but it is possible that none will be found. I suggested that this gap arises from certain innate drives toward retributive punishment, and a mismatch between these innate drives and what is deemed normatively appropriate. Third, and finally, I have

argued that this retributive gap has three potentially significant social implications: (i) it could lead to an increased risk of moral scapegoating; (ii) it could erode confidence in the rule of law; and (iii) it could present a strategic opening for those who favour non-retributive approaches to crime and punishment.

Bibliography

Alexander, L and Ferzan, K. (2009) *Crime and Culpability: a Theory of Criminal Law*. Cambridge: Cambridge University Press.

Asaro, P. (2011). A Body to Kick and Still no Soul to Damn: Legal Perspectives on Robotics. In Lin, P., Abney and Bekey (eds) *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

Atran, S. (2002). *In Gods we Trust: The Evolutionary Landscape of Religion*. Oxford: OUP.

Barrett, J. (2004). *Why would anyone believe in God?* Walnut Creek, CA: AltaMira Press.

Boonin, D. (2008). *The Problem of Punishment*. Cambridge: Cambridge University Press.

Bostrom, N. (2014) *Superintelligence: Paths, Strategies and Dangers*. Oxford: OUP

Bostrom, N. (2012) The Superintelligent Will: Motivation and Instrumental Rationality in Artificial Agents. *Minds and Machines* 22(2): 71-85

Boyer, P. (2002). *Religion Explained*. London: Vintage.

Brahman, Kahan and Hoffman (2010). Some Realism about Punishment Naturalism. *University of Chicago Law Review* 77: 1531.

Brown, D. O. (1991). *Human Universals*. New York: McGraw-Hill

Bryant, C. (2015). Worker killed in Volkswagen Robot Accident. *Financial Times* 1 July 2015.

Calo, R. (2015) Robotics and the Lessons of Cyberlaw. *California Law Review* 103(3): 513-563.

Calo, R., Kerr, I and Froomkin, M. (2016) (eds). *Robot Law*. Edward Elgar Publishing

Carlsmith and Darley (2008). Psychological Aspects of Retributive Justice. In Zanna (ed) *Advances in Experimental Social Psychology*. San Diego, CA: Elsevier.

Chisan Hew, P. (2014). Artificial moral agents are infeasible with existing technologies. *Ethics and Information Technology* 16: 197-206.

Duff, R.A. (2007). *Answering for Crime: Responsibility and Liability in Criminal Law*. Oxford: Hart Publishing.

Ford, M (2015). *The Rise of the Robots*. London: Oneworld Publications.

Gintis, H. (2011). *The Bounds of Reason*. Princeton, NJ: Princeton University Press.

Gunkel, D. (2012). *The Machine Question*. Cambridge, MA: MIT Press.

Hart, HLA (1968). *Punishment and Responsibility*. Oxford: Clarendon Press.

Hellstrom, T. (2013). On the moral responsibility of military drones. *Ethics and Information Technology* 15: 99-107

Hinds, P., Roberts, T., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction* 19: 151–181.

Jensen (2010). Punishment and Spite: The Dark Side of Cooperation. *Philosophical Transactions of the Royal Society B* 365: 2635-2650

Kaplan, J. (2015) *Humans Need Not Apply*. Yale University Press.

Kim, T., & Hinds, P. J. (2006). Who should i blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *Proceedings of RO-MAN'06*. Available at http://web.stanford.edu/~phinds/PDFs/Kim-Hinds-ROMAN.pdf (accessed 22/11/15).

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication and Science*. DOI: 10.1080/1369118X.2016.1154087

Kiesler, Sara, and Jennifer Goetz. (2002). Mental Models of Robotic Assistants. In *Conference Proceedings of CHI 2002, Extended Abstracts on Human Factors in Computing Systems*: 576-77.

Kramer, M. (2011). *The Ethics of Capital Punishment*. Oxford: OUP.

Levy, N. (2014). *Consciousness and Moral Responsibility.* Oxford: OUP

List, C and Pettit, P (2011). *Group Agency: The Possibility, Status and Design of Corporate Agents*. Oxford: OUP

MacDorman, K F, & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*: 7(3): 297–337

MacDorman, Karl F. (2006). Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. *Proceedings of the ICCS/CogSci-2006:Toward Social Mechanisms of Android Science*.

MacDorman, Karl F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior* 25(3): 695– 710

Marin, A.L., Doori Jo, and Sukhan Lee. (2013). Designing Robotic Avatars. Are User's Impressions Affected by Avatar's Age? *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI 2013, March 3-6.

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology* 6(3): 175-183.

Matthias, A. (2011). Algorithmic moral control of war robots: Philosophical questions. *Law, Innovation and Technology* 3(2): 279-301.

Moore, M. (1993). Justifying Retributivism. *Israel Law Review* 27:15.

Moore, M. (1997). *Placing Blame: A General Theory of Criminal Law*. Oxford: OUP.

Muehlhauser, L and Helm, L. (2013). The Singularity and Machine Ethics. In Eden, A., Moor, J., Soraker, J and Steinhart, E. (eds) Singularity Hypotheses: A Scientific and Philosophical Assessment. Dordrecht: Springer.

Prevost, A.M. (1992). Race and War Crimes: The 1945 War–Crimes Trial of General Tomoyuki Yamashita. *Human Rights Quarterly* 14: 303.

Pinker, S. (2011). *The Better Angels of Our Nature*. London: Penguin.

Purves, D., Jenkins, R. and Strawser, B. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethical Theory and Moral Practice* 18: 851-872.

Robinson, P.H. (2013). *Intuitions of Justice and the Utility of Desert*. Oxford: OUP.

Robinson, P.H. and Kurzban, R. (2007). Concordance and Conflict in Intuitions of Justice. *Minnesota Law Review* 91: 1829, 1892 (2007);

Robinson, P.H. Kurzban, R. and Jones, O.D. (2007). The Origins of Shared Intuitions of Justice. *Vanderbilt Law Review* 60: 1633.

Sainato, M. (2015). Stephen Hawking, Elon Musk and Bill Gates Warn about Artificial Intelligence. *The Observer* 19 August 2015.

Scherer, M. (2016). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies. *Harvard Journal of Law and Technology*. Forthcoming.

Simpson, T. and Muller, V. (2016). Just Wars and Robots' Killings. *The Philosophical Quarterly* DOI: 10.1093/pq/pqv075.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy* 24(1): 62.

Vincent, N. (2011). A Structured Taxonomy of Responsibility Concepts. In N Vincent, I van de Poel & J van den Hoven (eds) *Moral Responsibility: Beyond Free Will and Determinism*. Dordrecht: Springer.

Walker, A. (2016). Why Self Driving Cars Should Never Have Steering Wheels. *Gizmodo* 24 February 2016 – available at http://gizmodo.com/why-self-driving-cars-really-shouldnt-ever-have-steerin-1758292942

Yamashita, In re (1946) 321 US 1.

Zimmerman, M. (2011). *The Immorality of Punishment*. Broadview Press.