VIRGINIA JOURNAL OF LAW & TECHNOLOGY

FALL 2016 UNIVERSITY OF VIRGINIA VOL. 20, NO. 02

The Co-Evolution of Autonomous Machines and Legal Responsibility

MARK A. CHINEN†

† Professor, Seattle University School of Law. My thanks to Joshua Fensterbush for his valuable research assistance and to my colleagues at the Seattle University School of Law for their helpful comments to earlier drafts. All errors remain my own. A book based on this article is expected to be published by Edward Elgar Publishing.

ABSTRACT

This Article sets out a possible trajectory for the co-evolution of legal responsibility and autonomous machines. Commentators have responded to the problem of legal responsibility for harms caused by such machines with already-existing legal doctrines related to defective products, agency law, and international humanitarian law, among others. There is a debate about the extent to which those doctrines in their current forms can address adequately the situations that will arise when autonomous machines become more prevalent. To the extent they do not, it is because of the law's general discomfort with associative responsibility, a discomfort shared and informed by most of the literature on ethics. The ethical literature most relevant to the problems of associative responsibility provides some guidance on the issue but no completely satisfactory answers. In turn, the concern that there will be gaps in responsibility for harms caused by machines leads to two interweaving lines of development. The first is to refine the concept of responsibility as a way to lessen that gap. The second is to reduce harm by designing autonomous machines with prosocial behaviors. If that second effort is successful, that very success, together with calls to grant legal personhood to machines for legal and pragmatic reasons and the human tendency to anthropomorphize, will strengthen what are now nascent calls to treat such machines as moral agents. This trajectory, however, must be placed in the context of society's current attitudes about how far responsibility in general should extend.

TABLE OF CONTENTS

I. INTRODUCTION

Autonomous machines such as self-driving automobiles and autonomous weapons systems are no longer a distant prospect, and the issue of how law can be used to prevent them from doing harm and how to assign responsibility if they do is more pressing.[1] This Article plots a trajectory for the evolution of legal responsibility and decision-making machines and systems.[2] At present, we address the issue with already-

[1] A recent article from the popular press is Adrienne LaFrance, *Can Google's Driverless Car Project Survive a Fatal Accident?* ATLANTIC (Mar. 1, 2016), http://www.theatlantic.com/technology/archive/2016/03/google-self-driving-car-crash/471678/http://www.theatlantic.com/technology/archive/2016/03/google-self-driving-car-crash/471678/. The first death in a self-driving vehicle occurred on May 7, 2016. *See* Bill Vlasic & Neal E. Boudette, *As U.S. Investigates Fatal Tesla Crash, Company Defends Autopilot System*, N.Y. TIMES (July 16, 2016), http://www.nytimes.com/2016/07/13/business/tesla-autopilot-fatal-crash-investigation.html?_r=0.

[2] The Article will assume there will be no upper bound on the sophistication of such machines. Municipal or international law could limit their development, but their perceived advantages and the diffuse nature of the threats they pose to most individuals and societies means such limits are unlikely to be imposed in the near term. Concerns about negative impacts of autonomous machines are raised most often with regard to

existing legal doctrines and principles, but the increasing ability of machines to decide for themselves is leading to the coevolution of legal norms and of the machines in question. Such developments are taking place along two parameters and lines of development. One parameter is the nexus between the machine and human activity. The legal system is trying as much as possible to associate the actions of autonomous machines and their consequences to individuals or groups of human beings, and the doctrines used include individual liability for human individuals, products liability, agency, joint criminal enterprise, aiding and abetting, conspiracy, and command responsibility. With modifications, such doctrines would seem to work relatively well for less sophisticated machines and more or less so in cases where sophisticated machines are clearly carrying out the will of human beings.

However, this is where the second parameter, the degree of autonomy of the machine as decision-maker, comes into play. The more autonomy machines achieve, the more tenuous becomes the strategy of attributing and distributing legal responsibility for their behavior to human beings. To be sure, there are strict liability doctrines, but in general, the law is more comfortable with assigning legal liability to someone when he is personally culpable for a harm and far less so with liability or guilt by association. In this sense, the law corresponds to prevailing views of moral responsibility. As machines become more sophisticated, their actions become less tied to human beings, and the assignment of legal responsibility to humans for what machines cause becomes less defensible. Thus, the strengths and weaknesses of specific proposals such as "use principles of products liability and other tort doctrines if using nanotechnology in medical treatment harms a patient" or "apply the principles of command responsibility if an autonomous weapon 'commits' an act that would constitute a war crime if a human committed it" depend in part on our comfort with the ‘solutions’ to the problem of associational responsibility. Even in cases where such

---

autonomous weapons. *See e.g.*, Markus Wagner, *The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapons Systems*, 47 VAND. J. TRANSNAT’L L. 1371 (2014) (discussing the negative impacts of autonomous weapons on existing law, ethics and politics); Human Rights Watch, *Losing Humanity: The Case Against Killer Robots*, HUMAN RIGHTS WATCH (Nov. 19, 2012), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots (arguing for a ban on autonomous weapons).

machines are clearly being employed by human beings, some of the incentives created by legal rules, such as the incentive to take due care, weaken because humans will be less able to supervise truly autonomous machines. And since at this point a machine cannot 'feel' legal sanctions, other purposes of the law, particularly those that motivate criminal law, are thwarted. As a result, we are forced to become more comfortable with group legal responsibility or responsibility by association, or face the prospect of manufacturers, owners or users of such machines becoming increasingly insulated from the law.

For these reasons, one would expect two things to occur. First, greater awareness of the permeability of responsibility could make associational responsibility more acceptable, and alternative forms of redress or compensation for harm, such as insurance, might be more emphasized. Second, some scholars urge that autonomous machines be given legal personhood to satisfy third parties who have been harmed by them while at the same time avoiding some of the problems raised by associative responsibility. In the future, we would expect to see designers try to instill a sense of legal responsibility within the machine itself. Of course, since as just discussed, machines are not cognizant of the law, far less do they 'appreciate' or 'value' it, all we can do is program machines to act as much as possible in conformity to the law, for example, by instructing autonomous cars to obey traffic laws or an autonomous weapon to obey the law of war. Of course this development is possible only to a certain extent: law cannot always be reduced to rules of decision. Besides, many of the legal issues involving autonomous machines will be retrospective in nature: we will need to determine *ex post* whether a machine's action has legal significance. Things will vary according to the level of sophistication of the machine or system, but over the long term, machines at the highest level of autonomy will need to be programmed in a way so that they are 'motivated' to engage in the kinds of prosocial behaviors the law is designed to promote. Of course, the case of HAL in *2001: A Space Odyssey* and critiques of Asimov's laws of robotics[3] show this can

[3] The laws figured as part of Asimov's science fiction *Robot* series. They are:

> A robot may not injure a human being or, through inaction, allow a human being to come to harm.
> A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

succeed to some extent, but as machines gain those kinds of prosocial capacities, it will strengthen calls already being sounded to grant autonomous machines legal and moral rights.

My arguments are set out in four parts. Part II surveys briefly the issue of autonomous machines and the existing legal approaches that frame and address the problem. Ultimately, law will need to address large and complex systems of humans and machines who work together. However, the law focuses primarily on individual responsibility, which dovetails with generally accepted understandings of moral responsibility. Applications of the law to groups still tend to frame the analysis in individualistic terms. This raises the question whether an approach designed with the individual in mind is well-suited to address large systems, because the knottiest issues involving humans and machines will raise problems of associational responsibility. Part III discusses the literature of group responsibility because many of its themes apply to issues of associational responsibility as applied to humans and machines. While that literature suggests some ways to address the problem of associational responsibility, ultimately it underlines how difficult the issue is. Part IV thus discusses the other route being considered: to instill legal and moral responsibility in the machines themselves. Part V concludes by putting these matters into a larger context: the extent to which autonomous machines will impact our understanding of responsibility will depend on a choice whether to allow the lines of responsibility to penetrate complex systems. Throughout, I refer to the literature on the moral responsibility of autonomous machines because the discussion of machines and responsibility seems best developed there.

> A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

ISAAC ASIMOV, I, ROBOT 42 (Gnome Press ed. 1950). For an evaluation of the three laws *see e.g.*, Keith Abney, *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 35, 42–44 (Patrick Lin et. al., eds. 2012).

II. AUTONOMOUS MACHINES AND LEGAL RESPONSIBILITY

A. The Issue

In a recent article, Peter Asaro sets out the challenges that autonomous machines pose to moral theory:

> [T]he crucial things we need are theories of punishment, agency and responsibility that apply to . . . complicated systems, systems of humans and machines working together. . . . Theories in which responsibility and agency can be meaningfully designed and shared, so that large organizations of people and machines can produce desirable results and be held accountable and reformed when they fail to do so.[4]

Asaro's challenges are interesting in several respects. First, for the most part, they appear instrumental in nature. It is a given for Asaro that autonomous machines will be part of everyday life; hence the need for theories of punishment, agency and responsibility that will ensure desirable results from the interaction of humans and machines and reform when needed. Second, what will eventually have to be addressed are not individuals primarily, but large, complicated systems or organizations instead. This reflects a growing reality in which the machines and systems in question are designed and manufactured by large organizations or through long supply chains in which sophisticated machines are already being used and in which such new machines will operate in systems or organizations of which people are also a part. Third, Asaro appears to assume machines will reach levels of autonomy at which it is as appropriate, or perhaps more so, to refer to "humans and machines working together" and "large organizations of people and machines," as it is to refer to "humans using machines in their work" or "large organizations of people who use machines." Asaro poses these challenges to the field of ethics, but they serve as a way of assessing how well law meets analogous challenges, and if not, how law might be changed to do so.

---

[4] Peter M. Asaro, *Determinism, Machine Agency, and Responsibility*, 2 POLITICA & SOCIETÀ 265, 291–92 (2014).

The way in which law and ethics are being used to address these challenges follows a pattern of cultural change and development proposed by J. M. Balkin.[5] Balkin argues all aspects of human culture, ranging from technology to the concepts comprising law and ethics have tool-like characteristics. Such tools have several features. Cultural tools are cumulative in that we apply already-existing tools to address new situations, and they have multiple uses.[6] Further, some of the multiple uses are unforeseen, leading to unexpected consequences, so that cultural tools take on a life of their own.[7] Finally, cultural tools are recursive: their use leads to new cultural realities that in turn require the tools involved to be modified to respond to those realities.[8] It follows from these features that cultural tools interact with other tools with the same effects of multiple uses in different contexts, unintended uses, the creation of new cultural situations, and recursion.

If Balkin's framework accurately depicts the development of cultural tools, we would expect societies to approach problems raised by autonomous machines by using preexisting legal and moral concepts and doctrines, but we would also expect the application of those tools to lead to unintended consequences that will in turn lead to modifications of those concepts of responsibility. The interaction between cultural tools does not take place in linear fashion, but even now, before the most sophisticated machines and systems have been manufactured and deployed, the literature is mapping out lines of development that fit into Balkin's framework. There is an interaction between the development of autonomous machines and the current systems of liability. Some observers have argued technology is running ahead of the law in this area, creating new facts on the ground to which law must respond.[9] However, it is more accurate to say that designers, programmers, policymakers, and jurists use the current systems of liability and the assumptions that underlie them, to frame and address potential legal issues raised by autonomous machines. In a sense, existing law is permitting and guiding their development. At the same time, observers are debating the extent to which this law is sufficient to address

[5] J.M. BALKIN, CULTURAL SOFTWARE (1998) (see in particular ch. 2).
[6] *Id.* at 32.
[7] *Id.*
[8] *Id.*
[9] UGO PAGALLO, THE LAWS OF ROBOTS 19–20 (2013).

foreseeable issues. This is leading to speculation about how the law will need to change and about how future machines will need to be designed.

B. Existing Legal Doctrines

Over the past five years, legal scholars have engaged in relatively detailed applications of current legal doctrines to problems that could arise with autonomous machines in the areas of tort, contract, and the law of war.[10] The situations being considered fall into three broad categories. First, a self-driving vehicle collides with a human and harms him. Second, a computer program operated by an online business enters into a contract with a human being where the online business did not authorize the contract. Third, an autonomous weapons system capable of selecting its own targets fails to distinguish between civilians and military personnel. Legal assessments of harms caused by self-diving vehicles gravitate towards products liability as the likely legal basis for assigning responsibility, with some discussion of agency law. Agency law is also the lens through which electronic contracting is assessed. Finally, the existing doctrines of command responsibility, and sometimes state responsibility, are applied to harms caused to civilians by autonomous weapons and systems. This subsection describes briefly how these areas of the law are being applied.

1. Cars, Contracts, and Weapons

Products. There is a growing literature on liability that could arise from autonomous vehicles. Gary Marchant and Rachel Lindor point out since driver error, the major cause of vehicle accidents, will be largely factored out, liability will focus on the manufacturer and others involved in the design of the vehicle or those involved with the infrastructure to support it.[11] Accidents that involve self-driving cars

---

[10] There are several commentators who have outlined the contours of a law of autonomous machines in the areas of the products liability, crime, contracting, and the law of war. In particular, *see* SAMIR CHOPRA & LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS chs. 2, 4 (2011); PAGALLO, *supra* note 9 at chs. 3–5.

[11] Gary E. Marchant & Rachel A. Lindor, *The Coming Collision Between Autonomous Vehicles and the Liability System*, 52 SANTA CLARA L. REV. 1321, 1327 (2012). For a recent discussion of the impact this will have on the automobile insurance industry, *see* Leslie Scism, *Driverless Cars Threaten to Crash Insurers' Earnings*, WALL ST. J.

will thus likely be assessed through products liability law concepts of design and manufacturing defects and adequate warning and instruction.[12] Some argue products liability law is already capable of addressing accidents caused by autonomous vehicles.[13] To the extent such concepts become unworkable, some propose strict liability be used to distribute costs among manufacturers, computer programmers, and engineers and to enhance insurance schemes.[14]

It might also be possible to view self-driving cars through an agency law framework. A self-driving car is not dissimilar to a human chauffer. If the car causes harm while it is transporting its owner, as principal, the owner of the car would be responsible since the car would have caused the damage while within the scope of its agency.[15] To the extent a frolic and detour would relieve an owner/passenger from liability, it would be possible to turn again to the manufacturer as designing a car capable of acting outside of the scope of its agency authority.

(July 26, 2016), http://www.wsj.com/articles/driverless-cars-threaten-to-crash-insurers-earnings-1469542958.

[12] Under current products liability law, liability can be found if there are defects in the design or manufacture of a product or in warnings about the product. Under U.S. law there are in general two tests whether a design is defective. The first is the consumer expectations test: "The article sold must be dangerous to an extent beyond that which would be contemplated by the ordinary consumer who purchases it, with the ordinary knowledge to the community as to its characteristics." RESTATEMENT (SECOND) OF TORTS § 402A cmt. i (1965). Under a cost-benefit approach articulated by the Third Restatement, a design is defective "when the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design . . . and the omission of the alternative design renders the product not reasonably safe[.]" RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2(b) (1998).

[13] Andrea Bertolini argues that current law should be able to adequately address issues raised by autonomous machines. Andrea Bertolini, *Robots as Products: The Case for Realistic Analysis of Robotic Applications and Liability Rules*, 5 LAW INNOVATION & TECH. 214, 222–23 (2007). *See also* JAMES M. ANDERSON ET. AL., RAND CORP., AUTONOMOUS VEHICLE TECHNOLOGY: A GUIDE FOR POLICYMAKERS 118–27 (2014) (applying existing products liability rules to autonomous vehicles); David C. Vladek, *Machines without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 132–40 (2014) (same).

[14] Vladek, *supra* note 13, at 146.

[15] Chopra and White take this approach, although this is supplemented by arguing that autonomous machines should be given some form of legal agency. *See* CHOPRA & WHITE, *supra* note 10, at 127–35.

Others, however, are less sure existing law will adequately compensate persons injured by autonomous machines. For example, Asaro agrees many of the legal issues raised by such machines will be covered by products liability rules[16] but fears it will be hard to tell whether a manufacturer has taken proper care in the design of the machine.[17] Samir Chopra and Laurence White are even less optimistic, particularly with regard to autonomous systems that are primarily computer driven. In their view, catastrophic damage caused by systems embedded in a tangible medium are most likely to lead to recovery under standard products liability rules.[18] Otherwise, they agree with Asaro that it will be hard for a plaintiff to meet the burden of showing that an artificial agent was defective, in part because it will be hard to show that there was a reasonable alternative design.[19] Further, since some machines and systems must be configured by the user, there will be arguments that the user has broken the chain of causation that would lead to liability of the manufacturer or designer.[20]

Contracts. With regard to contract, Chopra and White suggest agency law be used to govern issues raised by autonomous contracting. They note in modern shopping websites, the principal "cannot be said to have a preexisting 'intention' in respect of a particular contract that is 'communicated' to the user."[21] Instead, "in the case of a human principal, the principal has knowledge only of the rules the artificial agent applies."[22] In such situations, in Chopra and White's view, apparent authority could be used to determine whether a principal is liable for contracts that have completed by the autonomous system.[23]

---

[16] Peter M. Asaro, *A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 169, 170 (Patrick Lin et. al., eds. 2012) [hereinafter *Body to Kick*].

[17] *Body to Kick*, *supra* note 16, at 171.

[18] CHOPRA & WHITE, *supra* note 10, at 144.

[19] *Id.*

[20] *Id.* at 137, 144. On the other hand, Marchant and Lindor worry that under current law, manufacturers will be deterred from designing such cars and thus call for legislative protection or federal preemption to allow their development. *See* Marchant & Lindor, *supra* note 11, at 1330–35, 1337–39.

[21] CHOPRA & WHITE, *supra* note 10, at 36.

[22] *Id.*

[23] *Id.* at 44. Apparent authority is "the power held by an agent or other actor to affect a principal's legal relations with third parties when a third party reasonably believes the actor has authority to act on behalf of the principal and that belief is traceable to

However, they concede certain changes would need to be made to some existing law because under some sources of agency law, such as the Restatement Third of Agency, computer programs do not appear to qualify as agents.[24] Ugo Pagallo largely agrees agency concepts best govern electronic contracting, but he believes they will work less well in cases of massive economic loss caused, for example, by autonomous trading systems.[25]

Autonomous Weapons. Commentators on the liability of autonomous weapons systems take a similar approach of applying existing law. During armed conflict, humanitarian law requires states distinguish between combatants and civilians and use force proportionally, the extent necessary to respond to armed force or to achieve a military objective.[26] Current law centers on the responsibility

---

the principal's manifestation." RESTATEMENT (THIRD) OF AGENCY § 2.03 (2006). Chopra and White argue that apparent authority correctly allocates costs among the operator, agent, and user. *See* CHOPRA & WHITE, *supra* note 10, at 47–48. In the case of the website and check-out, the principal holds out the website as the vehicle through which the user is able to contract. In most situations, under apparent authority the principal would be bound by the actions of the electronic agent even though the principal is unaware of the precise details of the particular contract involved. However, costs would shift to the user if it is unreasonable for the user to believe such power exists, for example, when a computer error causes the program to offer a product at an unrealistically low price. *Id.*

[24] CHOPRA & WHITE, *supra* note 10, at 50. The Restatement Third, for example requires that an agent be a "person." A person is defined as "(a) an individual; (b) an organization or association that has legal capacity to possess rights and incur obligations; (c) a government, political subdivision, or instrumentality or entity created by a government; or (d) any other entity that has legal capacity to possess rights and incur obligations." RESTATEMENT (THIRD) OF AGENCY §§ 1.01, 1.04(5) (2006). Allowing a computer program or autonomous machine to serve as an agent, would require the law to confer on machines the legal capacity to possess rights and incur obligations.

[25] PAGALLO, *supra* note 9, at 99. Pagallo argues in the case of robot traders that the traditional view of treating a robot as the tool of human beings, and thus attributing responsibility only to humans, is problematic for three reasons. First, it seems inapt to describe sophisticated robots needed for large-scale trading as tools. Second, Pagallo points out that just because a human has delegated some authority to a robot, it does not necessarily follow that the human is responsible for the robot's actions. Third, the robots-as-tools approach does not help in the distribution of responsibility between human beings. *Id.* Elsewhere, Pagallo argues humans have a claim not to be financially ruined by the decisions of their robots. *Id.* at 102.

[26] For a discussion of the principles of the law of war, including the principles of distinction, proportionality, and military necessity, *see* OFFICE OF GENERAL COUNSEL,

of the individual soldier, the officer in the field, and the commanding officer. Several commentators believe current law is ill-suited to address a war crime 'committed' by an autonomous weapons system. The weapon itself could not be tried, and it is unclear whether an officer in the field, let alone the commanding officer, would have the mens rea required to cause him or her to be liable for a war crime 'committed' by an autonomous weapon.[27] However, this would depend on the circumstances. An officer that instructs a machine to commit a war crime would obviously be liable for that crime.[28] Further, just as some commentators contend strict liability should be used with self-driving cars, some argue that if a superior or commanding officers is not found liable, the state itself could be found responsible under the international law of state responsibility. The weapon's actions would be attributed to the state, since it was deployed as part of a state function.[29]

2. Laws Related to Groups

Scholars are thus divided in their assessments whether current law is able to resolve issues of liability that arise in tort, contract, and international law, and there is a sense in which to resolve the debate, much of this will need to be worked out through individual cases, legislation and other forms of governance.[30] However, one can go one

---

U.S. DEPARTMENT OF DEFENSE, DEPARTMENT OF DEFENSE WAR MANUAL 50–69 (June 2015).

[27] Jack M. Beard, *Autonomous Weapons and Human Responsibilities*, 45 GEO. J. INT'L L. 617, 651–57 (2015); Human Rights Watch, *Losing Humanity: The Case Against Killer Robots*, HUMAN RIGHTS WATCH (Nov. 19, 2012), https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots (arguing that it will be difficult to hold military commanders liable for war crimes committed by robots); Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 62, 70–71 (2007).

[28] In this regard, Christopher Toscano argues that the existing law of command responsibility should be enough to hold persons responsible for crimes caused by autonomous weapons. Christopher P. Toscano, "*Friend of Humans": An Argument for Developing Autonomous Weapon Systems*, 8 J. NAT'L SEC. L. & POL'Y 189, 235–37 (2015).

[29] See Marco Sassoli, *Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified*, 90 INT'L L. STUD. 308, 315–16 (2014).

[30] Ugo Pagallo places current scholarship along a spectrum ranging from those who argue autonomous machines will raise no novel issues of legal responsibility, to those who argue there will be new forms of responsibility but humans will remain

step further in this assessment of law. As discussed, Asaro's challenges are posed to larger systems. However, it is fair to say most legal and moral theories of responsibility use the individual as the starting point, and the doctrines and moral principles designed to address larger groups are ancillary to doctrines primarily addressed to the individual. Of course, law has always had to do with groups, and a number of legal doctrines attempt to address groups as such. Products liability law deals with large enterprises. There are doctrines of aiding and abetting and joint tortfeasorship. Laws that govern business entities regulate the components of large groups; subsets of agency, partnership, corporate, and limited liability law set out rights and responsibilities of the owners and managers of the firm. In criminal law there is conspiracy in some domestic legal systems and joint criminality in others and at the international level. By definition international law has to do with nation states. Finally, if Asaro is correct that we should be concerned with large systems of humans and robots, there is of course the whole of regulatory law in which legislation and underlying regulations address almost every aspect of modern societies.

Law therefore does treat large systems. At the same time, when theories of punishment, agency and responsibility are involved, law tends to become individualistic in nature, and responsibilities to others become understood as a set of binary relations, even though those theories are justified in part by their impacts on the larger society. In products liability, the manufacturer is of course liable for defective products, but the manufacturer is itself understood as a unitary whole in the analysis. In contract law, the focus is on two contract parties, with some doctrines that address the interests of third parties. Even corporations and other business entities are understood as individual actors in their relations with third-party creditors. In criminal law, the crime of conspiracy is controversial in some legal systems precisely because it does not sufficiently focus on individual culpability. The same is true for joint criminal enterprise. In the law of war, the analysis of war crimes focuses on the actions of individual soldiers and

---

responsible, and finally to those who argue new forms of legal responsibility will need to rest on the machines themselves. Ugo Pagallo, *What Robots Want: Autonomous Machines, Codes and New Frontiers of Legal Responsibility*, *in* HUMAN LAW AND COMPUTER LAW: COMPARATIVE PERSPECTIVES, 25 IUS GENTIUM 47, 53 (Mireille Hildebrandt & Jeanne Gaakeer, eds. 2013).

commanders. On the level of state responsibility in international law, the state is viewed as a monolithic whole, with little attention paid to the components of the state. Even in the area of regulation, when enforcement is involved, the subjects of enforcement tend to focus on individual subjects or business or political subjects viewed as individuals. The question becomes whether this emphasis on individual responsibility poses potential problems for legal systems of responsibility, agency, and punishment that will need to reach large systems of humans and machines.

C. The Moral Responsibility of the Individual

I save a final assessment of the law's ability to meet Asaro's challenge for the conclusion, but here, it is helpful to explore more fully current views of moral responsibility. The law's emphasis on individuals when responsibility is involved stems in large part from our views on ethics.

1. Major Approaches

Responsibility has been defined as "the quality or state of being responsible,"[31] and in turn, to be responsible has been defined in part as "liable to be called on to answer;" "liable to be called to account as the primary cause, motive, agent;" or "liable for legal review or in case of fault to penalties."[32] These definitions reflect various understandings of responsibility common in the West. Andrew Eshleman describes the field[33] as beginning with early Greek philosophers who wrestled with fatalism spurred by the gods' intervention in human affairs. Aristotle's major work, the *Nichomachean Ethics*, however, articulates the problem

---

[31] *Responsibility Definition*, MERRIAM-WEBSTER DICTIONARY (2015), http://www.merriam-webster.com/dictionary/responsibility.

[32] *Responsible Definition*, MERRIAM-WEBSTER DICTIONARY (2015), http://www.merriam-webster.com/dictionary/responsible. On the idea that responsibility entails providing an explanation for oneself, *see* Andreas Matthias, *The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, 6 ETHICS INFO. TECH. 175, 175 (2004) ("When we judge a person responsible for an action, we mean . . . that a person should be able to offer an explanation of her intentions and beliefs when asked to do so . . .").

[33] Andrew Eshleman, *Moral Responsibility*, in STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed. Summer 2014), *available at* http://plato.stanford.edu/entries/moral-responsibility/.

in ways still discussed when machines are involved. Aristotle conceives of responsibility as a person being subject to moral blame or praise for ones feelings or actions. That in part depends on whether such feelings or actions are voluntary or involuntary.[34] In this regard, Aristotle argues an action done through ignorance is a form of involuntary action and thus not subject to moral blame.[35] He continues by asserting that an action is praiseworthy if done through rational choice, chosen as a way to achieve an end that has been determined through deliberation.[36]

In Eshleman's view, Aristotle leaves unanswered a question still being debated: whether a person is subject to praise or blame because the individual in question herself has merited it, a merit-based view, or whether individuals are praised or blamed to influence their behavior, a consequentialist view.[37] This debate intertwines with another concerning scientific or theological determinism, the idea that all events are determined by the physical laws of the universe or by an omniscient and omnipotent God. Incompatibalists believe that if determinism is true, no one can be morally responsible because one's actions are not voluntary. In contrast, compatibalists argue a person can be morally responsible even if important aspects of one's identity and actions are determined outside of oneself.[38] Eshleman observes that merit-based views of responsibility tend towards incompatibalism, whereas consequentialists tend towards compatibilism. "[P]raising and blaming could still be an effective means of influencing another's behavior, even in a deterministic world."[39]

Peter Strawson tries to resolve these debates by shifting focus from the justifications for moral praise or blame to the practice of moral praise or blame itself.[40] Strawson argues that in our relationships, we demand some degree of goodwill or regard on the part of those who

---

[34] ARISTOTLE, NICOMACHEAN ETHICS, Book III, ch. 1, 37 (330 BCE) (Roger Crisp, trans & ed. rev. ed. 2014).
[35] *Id.* at 38–39.
[36] *Id.* at chs. 2–3, 40–44.
[37] Eshleman, *supra* note 33.
[38] *Id.*
[39] *Id.*
[40] Peter F. Strawson, *Freedom and Resentment*, 48 PROC. BRIT. ACAD. 1 (1962), *reprinted in* PETER STRAWSON, FREEDOM AND RESENTMENT AND OTHER ESSAYS 1 (Routledge ed., 2008).

stand in relation to us, and we have certain reactive attitudes, such as gratitude or resentment, when that demand is either met or thwarted.[41] Sometimes those reactive attitudes can be suspended when a counterpart has an excuse, so his behavior is not a violation of the demand for goodwill, or when the person for some reason is not able to engage in everyday interpersonal relationships.[42] As Eshleman puts it, under this view, "[w]hereas judgments are true or false and thereby can generate the need for justification, the desire for good will and those attitudes generated by it possess no truth value themselves, thereby eliminating any need for an external justification."[43] Thus, one of Strawson's major contributions is to avoid metaphysical questions by looking at the community in which judgments are made, a community in which certain expectations about one's behavior towards one another have been adopted.[44] One implication of Strawson's approach is that any system of responsibility used to address the use of autonomous machines will necessarily be shaped by the communities in which moral judgments are made. As will be discussed below, this opens up space for communities, if they choose, to consider new forms of responsibility to accommodate machines.

In Eshelman's view, much of the contemporary literature has been devoted to responding to Strawson's contributions. Several strands are interesting for purposes of this Article. One is the distinction some scholars make between types of responsibility. Gary Watson, inspired by John Dewey, focuses on responsibility as a kind of self-disclosure: our actions express our commitments, morals, etc.[45] This self-disclosure leaves people open to moral appraisal for the various ends they choose. As Angela Smith puts it, a person is responsible for something because "she is connected to it in a way that it can, in

---

[41] *Id.* at 6–7.
[42] *Id.* at 7–10.
[43] Eshleman, *supra* note 33. For Strawson, such responsibility need not be justified for their consequentialist effects. "It is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes, in regulating behavior in ways considered desirable . . . . What *is* wrong is to forget that these practices, and their reception, the reactions to them, really *are* expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes." Strawson, *supra* note 40 at 27.
[44] *See* Philip Pettit & Michael Smith, *Freedom in Belief and Desire*, 93 J. PHIL. 429, 440–41 (1996).
[45] Gary Watson, *Two Faces of Responsibility*, 24 PHIL. TOPICS 227, 227–28 (1996).

principle, serve as a basis for moral appraisal of that person." [46] *Attribution* is used to refer to the connection between the person and the act: "Conduct can be attributable or imputable to an individual as agent and is open to appraisal that is therefore appraisal of the individual as adopter of ends."[47] This is distinct from *holding* someone responsible, usually in the negative sense of blaming that person for something, on the other.[48] Fairness issues arise here because holding someone responsible for something involves such negative consequences and entails the ability to make demands on that person. Hence, Watson argues "[i]t is unfair to impose sanctions upon people unless they have a reasonable opportunity to avoid incurring them."[49] One result of there being different kinds of responsibility is that it might be possible to choose or reconcile various issues that arise from 'harsher' forms of responsibility by making do with other, less problematic forms.

Finally, some commentators have focused on responsibility as requiring someone to give an account of her actions or attitudes. Marina Oshana is one of the proponents of this view. She writes, "[w]hen we say a person is morally responsible for something, we are essentially saying that a person did or caused some act (or exhibits some trait or character) for which it is fitting that she give an account."[50] This view presumes the individual in question meets some requirements of agency, has performed some act or exhibited a characteristic subject to certain moral standards, and has fallen short of those standards.[51] Finally, "the accountability interpretation assumes the actor possesses and is able to

---

[46] Angela Smith, *On Being Responsible and Holding Responsible*, 11 J. ETHICS 465, 465–66 (2007).
[47] Watson, *supra* note 45 at 229.
[48] As might be expected, various accounts can overlap. For example, R. Jay Wallace synthesizes Strawson's view of moral responsibility based on the reactive emotions and Kantian views of the responsibility based on individual autonomy to suggest it is reasonable to hold a person morally responsible (in the sense of subjecting that person to certain reactive emotions) if that person is capable of reflective self-control. R. JAY WALLACE, RESPONSIBILITY AND THE MORAL SENTIMENTS 160–61, 226 (1994).
[49] Watson, *supra* note 45 at 237 (emphasis omitted).
[50] Marina A.L. Oshana, *Ascriptions of Responsibility*, 34 AM. PHIL. QTY. 71, 77 (1997).
[51] *Id.*

exercise certain capacities, rationality, self-awareness, an ability to appreciate and reply to telling questions, and the like."[52]

Despite the differences among these accounts of moral responsibility, it is striking that each tends to focus on the individual and makes common assumptions about the persons who are the subject of moral assessments. The proponents of responsibility as answerability set out requirements that resonate with those required for Aristotelian moral praise or blameworthiness: it is fair to hold a person responsible for her actions if she is aware of the consequences of those acts and engages in them freely. This set of assumptions would also be consistent with certain criteria for the 'rules' of interpersonal reactions, because actions that justify gratitude and resentment depend in part on at least weak assumptions about the rationality, freedom, etc. of the persons involved as they live in relationship with each other. Further, scholars such as Watson and Smith agree that if one moves beyond attribution to holding someone responsible, some degree of freedom and control over one circumstances is necessary before sanctions are appropriate.

2. Implications for Autonomous Machines

Autonomous machines raise problems under all such versions of responsibility. If a machine is simply a tool, the subject of moral appraisal would obviously focus on the person who used it. A person who has no control over the actions of an autonomous machine would normally be absolved of responsibility, just as would a person who had no control over another person who committed a wrongful act. We tend to avoid holding a person responsible for the acts of another, even when there are links such as familial ties between people. Not only are there arguments that holding a person responsible for what someone else has done is unfair and unjustified under commonly-held views, it undermines much of the incentive power law exerts. If an individual believes she will be held responsible even if someone else primarily is, she will have little incentive to take care. Or if she is in a position to

[52] *Id.* These versions of responsibility can be mixed. David Shoemaker argues that a theory of ethics would encompass three understandings of responsibility: responsibility as attributability, answerability, and accountability. David Shoemaker, *Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility*, 121 ETHICS 602, 630–31 (2011).

prevent another person from doing harm, she has less incentive to do so, since she will be held responsible by association anyway. [53]

Under prevailing notions of responsibility, the issue for autonomous machines becomes if a truly autonomous machine can be said to be the primary cause of a particular harm, to what extent is it fair and appropriate to hold human individuals or groups responsible too? The concern that no one will be legally responsible has, as discussed earlier, caused some observers to revisit current understandings of associational responsibility. As Balkin would argue, it is natural to start with well-accepted doctrines that lend themselves to greater associational responsibility. Subpart B noted some observers want to expand strict liability in the area of self-driving cars (which would be applicable to other civilian uses of autonomous machines, such as medical applications of nanotechnology) and to state responsibility for autonomous weapons. Under strict liability, culpability is not taken into account; it is sufficient that there is some relevant association between the respondent and the harm, such as the owner of the land in the paradigmatic *Fletcher* case. Yet, commentators acknowledge this approach is problematic exactly because of strict liability's associational character. David Vladek, for example, argues strict liability should be applied to the manufacturer of self-driving cars,[54] but he concedes this approach can be unfair to the manufacturer. Hence, he suggests the law provide a way for manufacturers to seek contributions from suppliers and computer programmers through a form of common enterprise liability.[55]

---

[53] For example, Mark Reiff argues it is counterproductive to hold individuals responsible for collective action. If an individual believes he will be found liable for wrongdoing committed by someone else, he will have an incentive to engage in such wrongdoing and reap its benefits since he can no longer avoid punishment by refraining from the wrongful act. Mark R. Reiff, *Terrorism, Retribution, and Collective Responsibility*, 34 SOC. THEORY & PRAC. 209, 242 (2008).

From a law and economics perspective, the imposition of strict liability has the effect of reducing hazardous activity. If strict liability is imposed against a manufacturer, it will pass on the costs of liability to consumers, who on the margins will turn to a cheaper, less dangerous product. The Bridge, *Economic Analysis of Alternative Standards of Liability in Accident Law*, LEGAL THEORY: LAW AND ECONOMICS, https://cyber.harvard.edu/bridge/LawEconomics/neg-liab.htm.

[54] Vladek, *supra* note 13, at 146.

[55] *Id.* at 148–49.

> A common enterprise theory permits the law to impose joint liability without having to lay bare and grapple with the details of assigning every aspect of wrongdoing to one party or another; it is enough that in pursuit of a common aim the parties engaged in wrongdoing. That principle could be engrafted onto a new, strict liability regime to address the harms that may be visited on humans by intelligent autonomous machines when it is impossible or impracticable to assign fault to a specific person.[56]

Under this approach, it would be unnecessary to find direct links between a computer designer or a manufacturer and the autonomous machine that was more directly involved in an accident. Since each participant was part of a common enterprise, it is reasonable to distribute responsibility to each participant. This argument, however, is not uncontroversial because Vladek's joint liability approach extends the reach of associational liability even further.

Vladek's recommendations echo more comprehensive approaches. In 2010, a working group of scholars produced a set of five principles or rules governing moral responsibility for computing artifacts.[57] Rule 2 reads:

> The shared responsibility of computing artifacts is not a zero-sum game. The responsibility of an individual is not reduced simply because more people become involved in designing, developing, deploying or using the artifact. Instead, a person's responsibility includes being answerable for the behaviors of the artifact and for the artifact's effects after deployment, to the degree to which these effects are reasonably foreseeable by that person.[58]

A rule like this would be needed to reach the members of large groups of programmers and engineers who will contribute to the design and

---

[56] *Id.* at 149 (footnote omitted).
[57] Keith W. Miller, *Moral Responsibility for Computing Artifacts: "The Rules"*, 13 IT PROFESSIONAL 57, 57 (May/June 2011). Versions of the rules and commentary are *available at* https://edocs.uis.edu/kmill2/www/TheRules/.
[58] *Id.* at 58.

manufacture of intelligent machines and human members of groups who use them. Rules like this present some challenges, particularly given the realities of computer design. Wendell Wallach and Colin Allen write:

> Given the complexity of modern computers, engineers commonly discover that they cannot predict how a system will act in a new situation. Hundreds of engineers contribute to the design of each machine. Different companies, research centers and design teams work on individual hardware and software components that make up the final product. The modular design of a computer system can mean that no single person or group can fully grasp the way the system will interact with or respond to a complex flow of new inputs.[59]

In this passage, Wallach and Allen use the design process to show how difficult it is to say that any one designer could foresee what a computer driven device would do in the future. In addition to the fact that the design process above seems to undermine the possibility that the effects of such artifacts will be reasonably foreseeable, such a rule expands the scope of responsibility in controversial ways, as I discuss below.

3. Autonomy and Agency

Because law and moral theory emphasizes individual culpability, current law fits best when the 'actions' of machines can be closely associated with humans, either because the machine is so unsophisticated that it can be understood as merely a tool or, in the case of sophisticated machines, as acting on behalf of a human principal. With regard to the machine as tool, at this point, it is difficult to imagine a machine that truly acts on its own. Autonomous machines fall within a range of lesser or greater autonomy and pose corresponding challenges to legal responsibility. The less sophisticated a machine is, the more appropriate it is to focus on the individual human or group of humans who used it, and any harm caused by such a tool is readily attributable to its users. It is unproblematic to say "he damaged his neighbor's

[59] WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG 39 (2010).

bushes with the pruning shears," thus attributing responsibility for the damage to the person who wielded the tool.

Further, by their very nature, autonomous machines and systems are being developed for use by human beings. The more the actions of autonomous machines can be associated with humans, the easier it is for the existing law of products liability, agency law, joint criminal enterprise, aiding and abetting, conspiracy, and command responsibility to respond to harm caused by those machines. If a completely autonomous machine is designed for and used or directed by human beings to achieve a particular end, it seems relatively straightforward to distribute liability for harms caused by that machine to the human or collection of humans who are associated with it.

The problem of legal responsibility and autonomous machines is thus ameliorated to the extent even fully autonomous machines can be characterized as designed for and used by human beings; legal responsibility for harm caused by a such machine can eventually be distributed to a human individual or to a collection of human beings who employ them. However, that claim is not absolute: it can be argued that the sophistication of a machine does impact legal liability and stretches current conceptions of that liability. Put in terms of agency law, a completely autonomous machine would be capable of engaging in the frolic and detour alluded to earlier, an action not readily attributable to the human being that would be associated with it. Peter Sparrow, who is concerned with the moral responsibility of autonomous machines, writes as follows:

> [A]utonomy and moral responsibility go hand in hand. To say of an agent that they are autonomous is to say that their actions originate in them and reflect their ends. Furthermore, in a fully autonomous agent, these ends are ends that they have themselves, in some sense, chosen. Their ends result from the exercise of their capacity to reason on the basis of their own past experience. In both of these things, they are to be contrasted with an agent whose actions are determined, either by their own nature, or by the ends of others. Where an agent acts autonomously, then, it is not possible to hold anyone else responsible for its actions. In so far as the agent's actions were its own and stemmed from its own ends, others cannot be held responsible for them. Conversely, if we

> hold anyone else responsible for the actions of an agent, we must hold that, in relation to those acts at least, they were not autonomous.[60]

Sparrow is agnostic whether machines will ever achieve the highest levels of autonomy such that they are acting for themselves. However, he argues that the more autonomous those machines are, "the less it seems that those who program or design them, or those who order them into action, should be held responsible for their actions."[61]

Sparrow is concerned with who will be morally responsible for the actions of autonomous weapons systems, and his worry that no one will be responsible leads him to conclude it would be unethical to use them.[62] Others disagree with Sparrow,[63] but even if he is right as to the

---

[60] Sparrow, *supra* note 27, at 65–66.

[61] *Id.* at 66. Bertolini shares similar doubts that machines will achieve what she calls "strong autonomy." Bertolini, *supra* note 13, at 222–23. Andreas Matthias shares this concern. He argues

> [T]here is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough *control* over the machine's actions to be able to assume the responsibility for them.

Matthias, *supra* note 32, at 177.

[62] Sparrow, *supra* note 27, at 66. *See also* Noel E. Sharkey, *The Evitability of Autonomous Robot Warfare*, 94 INT'L REV. OF THE RED CROSS 787 (2012) (arguing that autonomous weapons should be banned because it will be difficult to hold human beings responsible for crimes caused by such weapons).

[63] For example, Kenneth Anderson and Matthew Waxman argue such reasoning seems particularly persuasive to those who have faith in ability of the laws of war and individual criminal liability to enforce compliance. They argue other mechanisms can be used to encourage such compliance and worry that holding individuals criminally liable for the use of autonomous weapons could have a chilling effect on the development of systems that might reduce harm to civilians. Kenneth Anderson & Matthew Waxman, *Law and Ethics for Robot Soldiers*, POL'Y REV., Dec. 2012 & Jan. 2013, at 35, 43. In this regard Toscano believes autonomous machines will be better than human beings at complying with the law of war while in combat. Toscano, *supra* note 28, at 224–42. In particular, he argues in the near term, since human beings will remain in the loop when autonomous weapons are used, existing civil and criminal liability mechanisms should be sufficient to address specific incidents involving such weapons. *Id.* at 235. Further, Toscano suggests such systems could actually enhance command responsibility because they constantly record data that could be used in investigations of any incidents. *Id.* at 238.

ethics of using such machines, his arguments are not necessarily applicable to legal responsibility. In an environment in which as a legal matter all things are permitted unless expressly prohibited, if a programmer, designer, or 'supervisor' of a machine cannot be held legally responsible for an autonomous machine's actions, it does not follow it would be illegal to program, design, or use it. Of course this exacerbates the issue because there might be machine-caused harms for which no one is legally accountable.

The lack of legal responsibility is worrisome for several reasons. From a purely instrumental perspective, one reason for developing autonomous machines is that they will achieve benefits human beings cannot realize alone. Eventually, self-driving cars will be safer than cars driven by humans, and although several observers argue strongly this will never be so, in theory autonomous weapons systems could eventually reduce the number of deaths caused in battle.[64] However, if designers, programmers, manufacturers, and officers are insulated from legal responsibility, the costs of harms caused by machines are shifted to consumers and civilians. Lack of such responsibility removes an incentive for designers, programmers, and manufacturers to avoid producing machines that pose an unreasonable risk. In the case of military applications, the failure to hold someone responsible could lead to impunity, with the result there would be little incentive to design machines and deploy them in ways that comply with the law of war.

The concern is law will find itself at an impasse. On the one hand, even machines that do not reach high levels of autonomy might still act in such a way that is hard under our current conceptions of legal responsibility to associate the machine's 'actions' with a human so that he or she could be held responsible legally for what the machine has done. On the other hand, such a machine is still without a "soul to be damned or a body to be kicked" so that it seems unsatisfactory and pointless to hold the machine responsible for itself. There appear to be two ways through the impasse. The first is to refine or redefine our understanding of associational responsibility. The second is to explore

---

[64] Toscano, *supra* note 28. Toscano argues autonomous weapons systems will be better than humans in reducing civilian casualties because they can remain objective, can act with greater caution, and can exceed human beings' biological limitations. *Id.* at 224–34.

the extent to which the machine itself can be deemed to bear legal rights and responsibilities. I consider each direction in turn.

III. ASSOCIATIONAL RESPONSIBILITY

A. The Literature of Group Responsibility

Rule 2 discussed above expands the responsibility of the designers, manufacturers, and users of autonomous machines. It is worth assessing such extensions under current ethical norms. The issues raised by associational liability are the subject of much of the literature of group responsibility. The field focuses generally on four interrelated problems.[65] The first relates to the question whether anyone beyond the human individual can be subject to responsibility. More precisely, the issue is whether a collective as such is capable of being subject to moral evaluation or whether, such an evaluation is really aimed at its members, since a group can only act through those members. Second, if in theory a collective can be subject to such judgment, are all groups susceptible to responsibility or are only certain kinds of collectives, such as corporations, morally answerable, while others, for example the crowd at a sporting event, are not? Third, when is it appropriate to distribute responsibility of a group to the members of the group? Finally, as a practical matter, even if a collective is morally responsible for some wrongful act and there are grounds for finding members in a collective responsible as well, what consequences should follow, particularly when those consequences will be felt by members, not the collective itself?

1. The Moral Responsibility of Groups as Such

All four questions have implications for the responsibility of autonomous machines, not only because in many instances, autonomous machines will be used in connection with groups, such as manufacturers, a supply chain, a military, or a government, but also because similar questions arise if the question involves a "group" of two: one human person and an autonomous machine working together

---

[65] Marion Smiley, *Collective Responsibility*, in STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Edward N. Zalta ed. Summer 2010), http://plato.stanford.edu/entries/collective-responsibility/.

causes harm. The question whether a group itself can be morally accountable involves an ontological or conceptual decision whether a group exists in and of itself as more than the sum of its parts or whether at base the group is shorthand for the actions of individual members.[66] In a sense, law has already answered this question: groups such as corporations and nation states are capable of incurring legal obligations and duties as such. However, the issue persists in other forms. The debate in corporate law between Adolf Berle and Gardiner Means on the one hand, who argue that the corporation should be understood as an entity of itself, and Michael Jensen and William Meckling on the other, who view the corporation as a nexus of contracts between and among its constituents, is one manifestation of the larger issue.[67] In some cases, the legal assumption that groups can be held legally responsible means that assigning responsibility to a group for what an

---

[66] For example, David Copp believes under some circumstances, a group can be found to be morally responsible for an action or outcome even though its members are not. David Copp, *The Collective Moral Autonomy Thesis*, 38 J. SOC. PHIL. 369 (2007). Sometimes the analysis turns on whether a group meets criteria for holding a human agent responsible. J. Angelo Corlett argues that some groups can be said to have an intention, act voluntarily, and have knowledge of the possible results of their actions so that the group can be morally responsible. J. Angelo Corlett, *Collective Moral Responsibility*, 32 J. SOC. PHIL. 573, 575 (2001). *See also* Philip Pettit, *Responsibility Incorporated*, 117 ETHICS 171 (2007) (arguing that certain groups meet criteria for being held morally responsible).

On the other hand, Colin Wight argues although the state is recognized as a legal subject, it is not in itself capable of independent action and should not be treated as a person for moral evaluation. Colin Wight, *State Agency: Social Action without Human Activity?*, 30 REV. INT'L STUD. 269, 278 (2004). For Wight, even though the state does have structures and causal powers that facilitate collective action, "such causal power that does emerge can only be accessed by individuals acting in cooperation with others." Colin Wight, *They Shoot Dead Horses Don't They? Locating Agency in the Agent-Structure Problematique*, 5 EUR. J. INT'L REL. 109, 128 (1999). *See also* John Hasnas, *Where is Felix Cohen When we Need Him?: Transcendental Nonsense and the Moral Responsibility of Corporations*, 19 J. L. & POL. 55 (2010) (arguing that the corporation cannot bear moral responsibility because it is a legal fiction); Pekka Mäkelä, *Collective Agents and Moral Responsibility*, 38 J. SOC. PHIL. 456 (2007) (arguing against collective responsibility).

[67] *See* ADOLF A. BERLE, JR. & GARDINER C. MEANS, THE MODERN CORPORATION AND PRIVATE PROPERTY 353–57 (1933); Michael C. Jensen & William H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, 3 J. FIN. ECON. 305 (1976). The question whether corporations can commit crimes is another example of this issue. For a discussion, *see* Edward B. Diskant, *Comparative Corporate Criminal Liability: Exploring the Uniquely American Doctrine Though Comparative Criminal Procedure*, 118 YALE L. J. 126 (2008).

autonomous machine does will not be a major leap in development, provided the machine can be said to belong to or is owned by the group. For example, it would not seem uncontroversial to hold a state responsible for harm caused by a robot weapon because such a weapon would be the property of the state. However, as discussed below, law's ability to hold groups legally responsible does not resolve all issues.

2. Types of Collectives

With regard to the question of what kind of collectives can be held morally responsible as such, several ethicists argue that long-lived groups with centralized decision-making systems and structures, such as an army or a corporation, can be subject to moral evaluation because those structures enable such groups to "think" and "plan" and to form and pursue goals, whereas more amorphous collectives like spectators at a sporting event or patrons in a restaurant should not be subject to moral evaluation. If some harm occurs at such an event or at the restaurant, individual spectators or customers will be evaluated, not the "group" itself. Other scholars propose an intermediate step that holds individuals subject to moral responsibility as members of teams. Ian Lee suggests

> [W]e think of a collectivity as being constituted and maintained by the self-identification of its members with the group. In this definition, the key concept is neither the group's identity nor its institutional features but the fact that its members regard themselves as the members of a collectivity. Collectivities are not quasi-persons, but *teams*.[68]

For Lee, a team exists when "its members regard themselves as the members of a team and adopt collectively rational principles as principles of action."[69] A team member can be held responsible for the actions of other team members because he or she has contributed to the goals of the team (thus having some degree of culpability), even though he or she was not directly involved in the harm caused by a team through one of his or her other teammates. Lee feels it is also appropriate to

---

[68] Ian B. Lee, *Corporate Criminal Responsibility as Team Member Responsibility*, 31 OXFORD J. LEGAL STUD. 755, 772 (2011).
[69] *Id.*

condemn the team as such over and above its members. He argues, "[c]ondemnation of the team draws attention to the contributory role that the team's norms played in producing the wrongdoing and to the responsibility of the member in relation to the content of those norms."[70] This condemnation is important because focusing only on the individual in effect absolves the team, with no impact on the team norms and structures that contributed to the harm.[71]

In the case of autonomous machines, principles like these would be useful in determining when it is appropriate to spread liability among a wider group of human 'participants,' such as among manufacturers, software programmers, and engineers in the case of autonomous vehicles, as Vladek proposes. Team concepts like those suggested by Lee would make it possible to hold looser associations of individuals or groups responsible for harms caused by autonomous machines. This conception would encompass the associations themselves and the members of the team. However, the issue of which kinds of groups can be subject to responsibility never completely disappears. Concepts like teams that allow looser affiliations of individuals to be held responsible simply raise the issue to another level. This is because whether there are enough coherent norms and structures in a collective to constitute a team will always be subject to debate. One can imagine for example some arguing that the contractual relations used to define the rights and duties of a production team allow us to characterize the loose affiliation as a team. Others however would argue that those 'norms' are too thin to create a common ethos and set of goals that one associates with sports teams.

3. The Distribution of Responsibility from a Group to its Members

As just discussed, the approaches taken by scholars of group moral responsibility to the first and second questions of whether and what kinds of groups might be candidates for moral evaluation and judgment are relevant to the responsibility of autonomous machines.

---

[70] *Id.* at 778.

[71] *See Id.* Amy Sepinwall also uses team ethics to argue it is appropriate for corporate officials to be held morally responsible for crimes committed by corporations. Amy J. Sepinwall, *Guilty by Proxy: Expanding the Boundaries of Responsibility in the Face of Corporate Crime*, 63 HASTINGS L. J. 411, 435–45 (2011).

However, the third and fourth questions appear to be the most germane to the issue of autonomous machines and associational responsibility. Recall that the third question is whether and under what circumstances the liability of a group can be distributed to its members. The literature tends to agree that since "judgments about the moral responsibility of [a collective's] members are not logically derivable from judgments about the moral responsibility of a collectivity,"[72] there must be some culpability on an individual's part before moral judgments about the collective can be transferred to her. Several grounds have been raised in this regard. Some ethicists argue that if a member shares the objectives of a group, it is appropriate she share responsibility for the group's actions to further them. As discussed earlier, shared goals are part of the basis for holding team members responsible for the actions of other teammates. Another approach focuses on shared benefits instead of shared goals: if a member benefits from the group it is fair that she share its burdens, including responsibility for harms committed by the group or other group members.

These approaches resonate with some of the directions the law has already taken, as discussed earlier. With regard to shared goals, it seems sensible that if manufacturers, software developers, and engineers share a common goal of producing an autonomous machine, it is not unfair to find them liable for harms caused by that machine. Similarly, since these people have benefited from the sale of such machines, it is appropriate they share any costs incurred by them. The shared goals and benefits approach does not necessarily require the individuals among whom responsibility is distributed be part of a particular kind of group, so long as there are goals and benefits common among the individuals involved. Shared goals and benefits also provide some of the moral underpinnings of agency law. Normally, the principal and agent share the same aims and both benefit from their relationship so that as a general matter, it seems appropriate that they share legal responsibility for the agent's actions. Thus, even though it is likely under current tort law that the manufacturer, and by extension, the software developer and the engineer, will be the primary focus of

---

[72] Virginia Held, *Can a Random Collection of Individuals be Morally Responsible?*, 67 J. PHIL. 471, 475 (1970).

attention, it could also be argued the owner/passenger of an autonomous vehicle should also bear some responsibility since he or she has employed the vehicle for his or her benefit.

At the same time, this last example reveals a limitation to the shared goal or benefit approach to associational responsibility. The approach makes several assumptions. One assumption is the very fact that sharing a goal becomes a predicate for responsibility. However, this situation is not always the case. It seems justified, for example, in the case of conspiracy, if everyone shares the same purpose of committing a crime to hold each member responsible for that crime (provided there is an actus reas). However, this assignment of responsibility is less obvious if the goal is not prohibited. The manufacturer, software developer, and engineer share the goal of creating an autonomous machine, but that goal is desirable. They certainly do not work together for the purpose of causing harm. If, however, they did not intend to cause harm, why should they be held responsible for that harm? Of course, it could be said tort law avoids the need to distinguish between appropriate and inappropriate goals by employing concepts of foreseeability: vague concepts like intent and the appropriateness of goals can be sidestepped because it is enough that people foresee their activities could cause harm. This approach is the impulse that informs Rule 2 discussed earlier.[73] However, we have now moved beyond a common goal approach.

Another issue with the common goal approach is it can ignore differences between a goal and the means to achieve it. Soldiers might share the same objective yet disagree about the means to fulfill a particular mission. Under current understandings of liability, that one solider commits a war crime is not imputed to fellow soldiers even though their primary goals are the same. It appears we must return to differentiating between legitimate and illegitimate ends and legitimate and illegitimate means,with the result that shared goals alone do not necessary justify associative responsibility.

[73] One of the issues raised by a foreseeability approach is since many things are foreseeable, the judgment that a risk was foreseeable is really a judgment about who should bear that risk.

Difficulties also arise if one uses shared benefits as a basis for associational responsibility. As discussed above, there is a sense in which people who share benefits from an activity should be responsible for costs incurred by it. However, as Reiff points out, benefits are not shared evenly among members of a group.[74] Further, he argues, receipt of a benefit is a different wrong than the original wrong.[75] These difficulties lead to two consequences. A benefit-based system of responsibility would require a method to apportion responsibility according to the amount of benefit received by a member. This apportionment might be possible in some cases but not in others. It might be that a benefit comes from a number of sources that cross group boundaries, thus making it hard to use a group-member, benefit-burden schema. Further, the difference between an original wrong and the benefit received raises several sub-issues. As Richard Vernon points out, one is the concern that any costs of sanctions for the original wrong will be disproportionate to any such benefit.[76] Moreover, it is not always the case that a person who receives benefits should share costs; for example, although citizens of a state certainly benefit from it, certain vulnerable individuals are protected with no expectation of return.[77] It thus appears that although there are certain justifications for distributing moral responsibility from group to member, no one justification is completely satisfactory.

4. The 'Pragmatics' of Group Responsibility

The question of consequences, the fourth major area of concern in the group responsibility literature, is conceptual and pragmatic. Much of this concern has been alluded to earlier. Whether a group itself should suffer consequences for a wrong depends in part on the purposes of moral sanctions and whether such consequences serve them. Sanctions are used for retribution, societal condemnation, or to deter future wrongs. Determining which purpose should be served and

---

[74] *See* Reiff, *supra* note 53, at 218–19.
[75] *Id.* at 219.
[76] Richard Vernon, *Punishing Collectives: States or Nations?*, *in* ACCOUNTABILITY FOR COLLECTIVE WRONGDOING 287, 300 (Tracy Isaacs & Richard Vernon eds., 2011) (citing Richard Vernon, *States of Risk: Should Cosmopolitans Favor Their Compatriots?*, 21 ETHICS AND INT'L AFF. 451, 451–69 (2007)).
[77] *Id.* (citing Robert Goodin, *What is So Special About Our Fellow-Countrymen?*, 98 ETHICS 663, 663–86 (1988).

whether a particular sanction will be effective in furthering it is hard enough in the case of individuals but becomes harder still when groups are involved: first, the group "has no soul to be damned and no body to be kicked," and second, those negative consequences often devolve to group members. The result can be seen as a balancing of the objectives of group sanctions with the fairness of distributing those sanctions downwards. It follows that the type of group sanction or the specific consequence might be relevant to whether they should devolve to the members of the group. For example, Avia Pasternak distinguishes between punishment and liability. She argues that it would be inappropriate to punish members for the wrongdoing of the group itself because for her, punishment is an expression of anger and moral judgment that should not be directed to individuals unless they are personally culpable.[78] However, Pasternak feels it is appropriate to distribute liability to members because liability does not carry the same sense of condemnation that punishment carries.[79] Although liability also imposes costs on members, it is not based on personal culpability;

---

[78] *See* Avia Pasternak, *The Distributive Effect of Collective Punishment*, *in* ACCOUNTABILITY FOR COLLECTIVE WRONGDOING 210, 212–16 (Richard Vernon & Tracy Isaacs eds., 2011). As Pasternak puts it, "[w]hen the group itself is the agent that behaves wrongly but its members are the ones who end up being condemned, then the necessary connection between responsibility and punishment . . . is broken." *Id.* at 216.

[79] *Id.* at 216–18. Pasternak's distinction between the respective bases for sanctions and liability raises the issue whether guilt by association under criminal law can be equated with being forced to pay the costs of state responsibility. It could be argued that they are distinct. Criminal sanctions can be severe and carry with them a strong sense of moral condemnation, hence the requirement for a particular mens rea and a heightened standard of proof. That distinction is not necessarily true for the commission of an internationally wrongful act. At the same time, the problems are analytically the same. Whether a state can commit a crime with the required mens rea, etc. is a subset of the question whether groups can be subject to responsibility. It is the same type of question as whether a state can commit a wrongful act. Assuming the answers to those questions are yes, the distributive questions are also similar. As discussed, we normally think that a person who bears the legal consequences of an act of another, whether criminal or not, must also be culpable to some extent. If a citizen bears criminal or civil sanctions for something committed by the state without such culpability, it is legal guilt or liability by association. International law could ground responsibility more on the fact that an injury has occurred and less on the fact a wrong has been committed. As is true in the area of transitional justice, this conception raises another kind of distributional problem: why citizens who are not responsible for a harm should be required to address it. *Id.* at 216 (citing Anthony Flew, *The Justification of Punishment*, 29 PHILOSOPHY 291, 293 (1954)).

, rather, it is based on the need to pay for the costs incurred when the law is broken and to compensate victims for harms. [80]

5. Summary

The issues addressed by the group responsibility literature resonate with the concerns of this Article. Part IV discusses technical and conceptual efforts to hold machines themselves liable for harms, but at this point in their development, autonomous machines resemble groups because they too have no souls or bodies that make them sensitive to moral condemnation or legal consequences. Earlier in this Article,[81] I discussed the concerns about impunity if no one is held responsible for harms caused by autonomous machines, creating the impulse to spread that responsibility. However, if we say a machine is primarily responsible for harm, to widen the circle of responsibility further to reach humans will mean distributing responsibility to others less culpable. As I have just argued, current justifications for the distribution of responsibility are never completely satisfactory. Either outcome, impunity on the one hand or responsibility by association on the other, seems undesirable.

B. Revisiting the Concept of Responsibility

Subpart A's review of the literature on group moral responsibility seems to jibe well with the law's current system of associational responsibility, but at the same time, it highlights some of the tensions within that system. Some commentators have tried to resolve these tensions by trying to alter responsibility in ways that better respond to the issues posed by autonomous machines.

1. A Shift in Emphasis to the Victim or Survivor of Harms or the Harm Itself

As discussed, the standard account of responsibility starts with the human individual, who sets goals for herself, acts freely to reach them, and is aware of the consequences of her actions. Departures from this standard view tend to take two directions. One direction is to shift

[80] *Id.* at 213.
[81] *See supra* text accompanying notes 60–64.

attention from the perpetrator of harm to the victim or survivor, or to the harm itself. This view draws from the moral principle that one should help someone who has been injured. Such a focus on the victim or survivor or on the harm itself has its merits. It justifies spreading costs among a wider group of people or throughout society as a whole without the need for any culpability of those asked to share the costs of the harm. Thus, one can imagine no-fault public or private insurance schemes that would compensate for damage to property or persons caused by autonomous machines. This option would have the benefit of pooling risk.[82] Similarly, as discussed earlier, a system of strict liability could spread liability costs among consumers that would cause them to choose less hazardous activities.

At the same time, this shift in emphasis raises other issues. Peter Singer has argued persuasively that a person who, without harm to himself, can assist another person has a moral duty to do so.[83] However, this claim is not uncontroversial. The fact that someone has been injured might serve as grounds for redress, but it does not fully answer why someone who has not caused the injury should provide it. Further, even if one accepts that an injury itself justifies a shared response, the question of how the costs of the injury should be shared remains, which raises its own issues of fairness. Pasternak argues in this regard there are three ways to distribute these costs: proportionally, equally, or randomly.[84] She points out distribution on a proportional basis is the most fair but sometimes hard to implement. A random distribution is the easiest to implement but the least fair. Therefore, an equal distribution of costs seems the most appropriate.[85] At the same time, even an equal distribution of costs requires some justification. In the case of the nation state, Pasternak suggests that citizens should accept an equal distribution of costs incurred when their government causes harm "because doing so is constitutive of a certain ethical understanding of the meaning of citizenship."[86] One can agree with Pasternak's view of citizenship or not, but this approach indicates that equal sharing does

---

[82] *See, e.g.*, Kenneth J. Meier & Robert M. La Follette, *The Policy Impact of No-Fault Automobile Insurance*, 6 POL'Y STUD. REV. 496, 502 (1987) (finding that no-fault insurance systems resulted in lower premiums to drivers).
[83] Peter Singer, *Famine, Affluence, and Morality*, 1 PHIL. & PUB. AFF. 229, 231 (1972).
[84] Pasternak, *supra* note 78, at 212.
[85] *Id.*
[86] *Id.*

not serve as its own justification; it is a compromise. Further, as is well known, insurance schemes tend to raise the problem of the moral hazard.[87] Finally, a no-fault system of compensation could reduce the benefits for victims that come from holding someone responsible for the harm.[88]

---

[87] *See* Kenneth Arrow, *Uncertainty and the Welfare Economics of Medical Care*, 53 AM. ECON. REV. 941, 961 (1963). The concern is that insurance will encourage people to take on more risk. On the moral implications of the moral hazard, *see, e.g.*, Will Braynen, *Moral Dimensions of Moral Hazards*, 26 UTILITAS 34 (2013); Rutger Claassen, *Financial Crisis and the Ethics of Moral Hazard*, 41 SOC. THEORY & PRAC. 527 (2015).

[88] In this regard, the literature suggests people who have access to compensation after an injury actually have worse health outcomes than those who do not have such access. Jason Thomson et al., *Attributions of Responsibility and Recovery Within a No-Fault Insurance Compensation System*, 59 REHABILITATION PSYCH. 247, 248 (2014) (*citing* Edward B. Blanchard et al., *Effects of Litigation Settlements on Posttraumatic Stress Symptoms in Motor Vehicle Accident Survivors*, 11 J. TRAUMATIC STRESS 337, 337–54 (1998); Belinda J. Gabbe et al., *The Relation Between Compensable Status and Long-term Patient Outcomes Following Orthopaedic Trauma*, 187 MED. J. AUSTL. 14, 14–17 (2007); and Ian Harris et al., *Association Between Compensation Status and Outcomes After Surgery: A Meta-Analysis*, 293 J. AM. MED. ASSOC. 1644, 1644–52 (2005)). In their study, Jason Thompson and his coauthors surveyed 934 road-trauma survivors to determine what variables might impact health outcomes in no-fault compensation systems "where access to compensation, medical and rehabilitation support is largely identical." *Id.* at 247. Their study finds that people in no-fault personal injury systems who feel others are responsible for their injuries have poorer post-accident outcomes than those who attribute responsibility to others. *Id.* at 247–48, 252. *Compare* Thomson et al., *supra*, *with* Michael Fitzharris et al., *The Relationship Between Perceived Crash Responsibility and Post-Crash Depression*, 49 PROC. ASSOC. AV. AUTOMOT. MED. 79 (2005) (finding that perceiving oneself as responsible for a crash is associated with higher rates of depression than when responsibility is seen to be shared, and to a lesser extent, when responsibility is attributed to another). Although Thomson et al. do not argue this conclusion, these results suggest that compensation alone is not sufficient to make injured parties whole, particularly if it is perceived that someone else is responsible for their injuries. It is unclear whether it would have made a difference to these people if the parties whom they blamed suffered some consequence for their actions. However, it might be that systems that focus more on the injury and less on fault will not be helpful to injured parties. Further, that injured people's health outcomes might be tied to attributions of responsibility underlines the importance of the responsibility problem when autonomous machines are involved. *But see* Toby Handfield, *Nozick, Prohibition, and No-Fault Motor Insurance*, 20 J. APPLIED PHIL. 201 (2003) (arguing on philosophical grounds there is no prima facie reason to believe the compensation afforded in a no-

2. Widening the Circle of Responsible Actors

Another alternative to an individualistic methodology for attributing responsibility retains the concept of culpability but reaches beyond the individual. In group responsibility, one approach is to cut the Gordian knot of distributional problems by emphasizing the group as the fundamental unit of concern. In a form of joint and several liability, each member is responsible for group wrongdoing as a matter of course, and each member's wrongdoing is attributed to each other member. "[W]hen one member of a community *commits* a wrong against a member of another, *all* members of the wrongdoer's community are equally responsible for that wrong, for each member of the community is an expression of its moral center."[89]

Other scholars suggest an intermediate step. F. Allan Hanson makes the case for extended agencies. He begins with the fundamental idea that moral responsibility for an act "lies with the subject that carried it out."[90] However, he points out that subjects are socially constructed and that in some circumstances, it seems more appropriate to view the subject as more than a human individual, particularly when technology is involved. He builds on an instinct that a person driving a car is in some sense different than the same person when she is riding a bicycle. "[I]f an action can be accomplished only with the collusion of a variety of human and nonhuman participants," he argues, "then the subject or agency that carries out the action cannot be limited to the human component but must consist of all of them."[91] Hanson then makes the case that an extended subject can be held morally responsible. First, like Watson and Smith (although he does not use their terminology), he distinguishes between a subject being responsible and a subject being held responsible and argues it is relatively straightforward to find extended subjects responsible for something in the former sense.[92]

---

fault scheme would be less adequate than that afforded by participation in a fault-based system).

[89] Reiff, *supra* note 53, at 227.

[90] F. Allan Hanson, *Beyond the Skin Bag: On the Moral Responsibility of Extended Agencies*, 11 ETHICS INFO. TECH. 91, 91 (2009). Bruno Latour makes similar arguments. *See* Bruno Latour, *On Technical Mediation*, 3 COMMON KNOWLEDGE 29 (1994).

[91] Hanson, *supra* note 90, at 92.

[92] *Id.* at 95.

Second, he asserts extended subjects can be seen as meeting at least some of the requirements normally required for moral responsibility for humans.[93]

With regard to awareness of the consequences and freedom of choice, Hanson agrees these conditions would need to be met to hold humans in an extended agency responsible, but points out awareness of consequences and freedom are necessary but insufficient conditions for responsibility. There must be some action as well, and humans often are unable to act without other parts of the extended agency: "[g]iven that moral responsibility cannot exist but for the action of the extended agency, it lies with the extended agency as a whole and should not be limited to any part of it."[94] He makes a similar case that extended agencies can be said to have intentions and argues an extended agency would be better at explaining causation than moral individualism; under some circumstances it seems much more plausible to say that an extended agency of humans and technology caused an event rather than the humans alone.[95] Finally, in Hanson's view, extended agency does a better job of explaining why a person's responsibilities increase when she moves from riding a bicycle, to driving a car, and then to being president of the United States with the codes to the nuclear arsenal.[96]

Given the difficulties we see in efforts to extend responsibility from machines to humans, it seems understandable why there have been other attempts to look beyond the human individual and to focus on extended subjects that include human beings and machines. Hansen's argument for extended agency is part of a strand of the philosophy of technology that posits the human person is being transformed by sophisticated technology that increases the human person's capacities in some ways hitherto not dreamed of and limits it in others.[97] Technology is becoming more sophisticated and more ubiquitous.

---

[93] Mark Coeckelbergh takes an analogous approach. *See* Mark Coeckelbergh, *Is Ethics of Robotics about Robots? Philosophy of Robotics Beyond Realism and Individualism*, 3 LAW INNOVATION & TECH. 241, 247 (2011).
[94] Hanson, *supra* note 90, at 96.
[95] *Id.* at 96–97.
[96] *Id.* at 97–98.
[97] "The person who surrenders her glasses, her telephone, her car, and her computer changes not only her instrumental abilities, but also her social life." BALKIN, *supra* note 5, at 24–25.

Perhaps parts of society will recognize that humans and machines are in a symbiotic relationship in which one cannot do without the other, so that we will become more comfortable with the idea that our subjectivities are part of a larger whole. If a machine with which I am associated causes harm, even though I did not personally intend that harm, I might not feel it unfair I be held responsible for it.

However, there are several challenges to this approach. One difficulty goes to the distinction discussed earlier between being responsible and being held responsible. It might seem fair to be considered responsible for harm in the sense that I and the machine constituted an extended agent that caused that harm. However, if we hold that extended agency responsible, then the fairness issues Watson identifies become relevant. The distributional issues that vex group responsibility arise again. An extended agency could be responsible for harm, but even if my subjectivity is so enmeshed in that agency that it is part of my identity, I, not the machine, will feel the negative consequences of being held responsible. I might view those consequences as unfair, particularly if I could not have reasonably foreseen that the machine involved would cause harm, a machine over which I ultimately had no control.

A second issue is analogous to the second question with which group responsibility wrestles: what kinds of groups can be subject to moral evaluation? Under an extended agency theory, an individual is part of a number of such agencies throughout the course of a day: when he steps in a car, when he sits at a computer terminal at work, whenever he 'uses' technology to perform a particular task. Thus, how closely tied to the agency must an individual be before he is considered part of it? Sometimes the extended agency that causes harm will be persistent, such as when a person regularly uses an autonomous car. At other times it will be almost ephemeral, even though the harm caused by such an agency is significant. That harm has occurred could by itself justify liability and distributing it among members of the extended agency. If, however, the agent is ephemeral, it will be tempting to fall back to more traditional, human-only agent analysis, even though under Hanson's framework, it was the extended agent that caused the harm.

The next Part discusses other attempts to redefine responsibility but in a different context: whether robots are deserving of moral concern. These efforts are all subject to the same criticism: the sophistication of autonomous machines and their ubiquity might lead

humans to fundamentally reconsider their moral frameworks and the role the human individual plays in ethics and in legal responsibility. However, in my view, such beliefs seem so fundamental that any changes will happen at the margins. If so, as discussed in Part II.C.3, it would seem Sparrow and other commentators are correct that there is a responsibility gap between our current understandings of responsibility and emerging technology and that such a gap is likely to persist.

IV. LEGAL RESPONSIBILITY AND MACHINE DESIGN

The exploration of how one might widen the scope of associational responsibility and the sense that a gap in responsibility is being created is accompanied by early forays into the legal responsibility of the autonomous machines themselves. At this point, such ideas seem farfetched. As discussed, machines are not cognizant of the law, far less do they 'appreciate' or 'value' it. However, as discussed below, engineers and commentators are giving serious thought to changing this point. The idea is to create machines with prosocial behaviors to minimize the possibility of harm and in the case of the most sophisticated of machines, to make machines themselves cognizant of their responsibilities to others, and to make them more susceptible to forms of punishment. This development is being done to make the machine itself more pliable and to make it more acceptable to human beings to hold the machine itself responsible when it causes harm and no recourse is available to another human being or organization. Other commentators have recommended giving serious thought to giving autonomous machines a kind of legal status, irrespective of whether they reach levels of true autonomy. This Part evaluates attempts to do so and their potential impacts.

A. Programming Law-Abiding and Ethical Machines

1. Rote Compliance with Law

At present, the most straightforward strategy is to program machines to act as much as possible in conformity to existing law, for example, by instructing autonomous cars to obey traffic laws or autonomous weapons to follow the laws of war. Here, existing products liability law, contract law, and the laws of war already impact machine design. However, programming machines to obey the law is possible only to a certain extent: law cannot always be reduced to a set of rules

of decision. For example, with regard to autonomous weapons, Patick Lin, George Bekey, and Keith Abney identify three reasons why what they term "operational morality" (essentially the doctrines of the law of war) alone will not insure compliance with the norms set out in the law.[98] First, weapons systems will become more autonomous. Second, intelligent systems will encounter complexities in the environment their designers could not anticipate, or they will be deployed in environments in which they were not intended to operate. Third, the technology itself will be complex, making it hard for systems engineers to predict how systems will behave when confronted with new information.[99] To build on these points, many of the legal issues involving autonomous machines will be retrospective in nature: we will need to determine whether something a machine has already done has legal significance. Ex ante programing will not always assist such ex post evaluations. Since programming a system to follow the law by rote will be insufficient in some instances, over the long term, the push will be for machines at the highest level of autonomy to be programmed so they are 'motivated' to engage in the kinds of prosocial behaviors the law is designed to promote.

2. The Debate on Autonomous Moral Agency

As an initial matter, it should be pointed out there is still a debate over whether it is possible to design moral machines. In a recent article for example, Patrick Hew argues artificial moral agents are infeasible given foreseeable technologies.[100] He begins with the Aristotelian view discussed earlier that a moral agent is one whose actions are subject to blame or praise and that such action is not morally blameworthy or praiseworthy unless it is voluntary.[101] Hew links these principles to a simple definition of intelligence: "anything that can *close a loop* from sensors to effectors without human intervention."[102] Hew points out a mouse trap would meet this definition of intelligence, yet we do not

---

[98] Patrick Lin et al., Ethics + Emerging Sciences Group, Autonomous Military Robotics: Risk, Ethics, and Design (2008), at 26, http://ethics.calpoly.edu/onr_report.pdf.
[99] *Id.*
[100] Patrick Chisan Hew, *Artificial Moral Agents are Infeasible with Foreseeable Technologies*, 16 ETHICS INFO. TECH. 197, 197 (2014).
[101] *Id.* at 199.
[102] *Id.* at 198.

view a mousetrap as a moral agent because, although it can close the loop between sensing and then trapping the mouse by itself, humans external to the mousetrap engineered its operative 'rules.' He then surveys technologies that are now being used in the area of artificial intelligence and argues that each one of them requires a human being outside of the intelligent system to supply its rules.[103] For Hew, this means actions taken by such systems are not voluntary; hence, they are not subject to moral blame or praise. Any blame or praise goes to the humans who supplied the rules for the system.

Hew's argument is reminiscent of the arguments related to physical or deistic determinism discussed in Part II and in a sense, is unanswerable. Earlier, I pointed out that the realities of modern software development and engineering would make it hard to attribute responsibility for machine-caused harm to any one person. Hew's point goes to another aspect of that issue: even if we use the corporation for whom the software developers and engineers work as the primary locus of responsibility, does it make a difference whether the system which has caused the harm is one generation removed from the corporate manufacturer or ten generations away? Hew believes it does not.[104]

Others, however, argue machines will reach such high enough levels of autonomy and sophistication that it will be hard to trace lines of responsibility back to a set of human beings. Andreas Matthias

---

[103] *Id.* at 198–200. These technologies are: self-replicating programs, self-modifying code, machine learning systems, self-regulating adaptive systems and meta-adaptive systems, self-organizing systems, and evolutionary computing. *Id.*

Hew concedes that "connectionist" approaches to artificial intelligence, such as neural networks, provide enough true autonomy to qualify an intelligent system as an agent whose actions are subject to blame or praise:

> [C]onnectionist systems are characterized by units interacting via weighted connections, where a unit's state is determined by inputs received from other units . . . . The opportunity is for unit states to define the rules used by other units. In this way, the connectionist system as a whole could come to supply its own rules.

Id. at 200. However, Hew believes that if the weights in a neural network system are provided by human beings, then the machine no longer qualifies as being autonomous. *Id.*

[104] Hew, *supra* note 100, at 201.

discusses three examples of how this is so. [105] In one example, an advanced Mars explorer is programmed to learn to avoid obstacles and navigate on its own by retaining images of terrain and information about how easy or hard it was to traverse so that the rover will act appropriately the next time it encounters similar terrain. Matthias argues if the rover falls into a hole, no one can be blamed: the operator on Earth did not give any manual controls and the programmer can point out the algorithm used was correctly implemented. The decision to move forward was based on facts about the planetary terrain that were encountered only after the rover had landed:

> The actual decisions of the control program were based not only on preprogrammed data, but on facts that were added to the machine's database only after it reached the surface of Mars: they are not part of the initial program, but constitute genuine experience acquired autonomously by the machine in the course of its operation.[106]

Matthias' illustration shows how the sophistication of machines could make it difficult to attribute accidents caused by the machine to humans. Lawrence Solum gives another example that forms a different response to Hew, which has to do with framing. In a well-known and prescient article written 20 years ago, Solum asks whether an electronic trustee could be given legal personhood.[107] He points out an electronic trustee could be designed to delegate certain decisions to a human trustee, which creates an argument that the human trustee is the 'real' trustee.[108] It follows that "the backup trustee must be the real trustee because there is a pragmatic need for discretionary decision making."[109] Solum, however, responds as follows:

> The objection that the AI is not the real trustee seems to rest on the possibility that a human backup will be needed. But it is also possible that an AI administering

---

[105] Matthias, *supra* note 32, at 176.
[106] *Id.*
[107] Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231, 1231 (1992).
[108] *Id.* at 1253.
[109] *Id.* at 1254.

> many thousands of trusts would need to turn over discretionary decisions to a natural person in only a few cases—perhaps none. What is the point of saying that in all of the thousands of trusts the AI handles by itself, the real trustee was some natural person on whom the AI would have called if a discretionary judgment had been required? Doesn't it seem strange to say that the real trustee is this unidentified natural person, who has had no contact with the trust? Isn't it more natural to say that the trustee was the AI, which holds title to the trust property, makes the investment decisions, writes the checks, and so forth? Even in the event that a human was substituted, I think that we would be inclined to say something like, "The AI was the trustee until June 7, then a human took over."[110]

One can take the illustration a step further. Suppose the electronic trustee commits a mistake that would be considered malpractice. Hew would conceivably hold the human trustee or the designers of the electronic trustee responsible for the wrong,[111] but if the electronic trustee has handled thousands to trusts by itself without mishap, it seems somewhat strained to hold the human trustee responsible the one time a mishap occurs.

3. Moral Machines, Susceptible to Punishment

Whether or not machines will be able to achieve 'true' intelligence and autonomy, there is a body of scholarship that is exploring how machines might be programmed to have prosocial behaviors and thus be more law-abiding.[112] For example, Lin, Bekey, and Abney believe it will be impossible to program autonomous weapons systems to always comply with the law of war. Consequently, such machines will need to be programmed to engage in rough forms of

---

[110] *Id.*

[111] Hew, *supra* note 100, at 201.

[112] In 2014, the U.S. Office of Naval Research offered a $7.5 million grant to a research team to develop robots to engage in moral reasoning. Nayef Al-Rodhan, *The Moral Code: How to Teach Robots Right and Wrong*, FOREIGN AFF. (Aug. 12, 2015), https://www.foreignaffairs.com/articles/2015-08-12/moral-code.

moral reasoning.[113] As Wendell Wallach and Colin Allen put it, "[m]oral agents monitor and regulate their behavior in light of the harms their actions may cause or the duties they neglect. Humans should expect nothing less of [autonomous moral agents]."[114] Both sets of authors recommend a hybrid approach, whereby machines will be given "top-down," deontological ethical rules, such as Asimov's three laws of robotics, or Kant's categorical imperative. The top-down approach has the advantage of providing rules that can apply in many situations but has the weakness of being too vague.[115] Thus, such machines should also be programmed to engage in "bottom-up" learning behaviors, whereby the rules of behavior will be able to evolve as machines are faced with specific situations.[116] For Wallach and Allen, the hybrid approach comes near to instilling a kind of virtue ethics in robots.[117]

Such approaches of course are enormously challenging and raise their own issues of responsibility. As Keith Abney points out, the attempt to program morality into robots highlights unanswered questions about competing ethical approaches.[118] The classic Trolley Problem first discussed by Philippa Foot has been cited as raising this problem. A trolley is running out of control down a track where five people are at work unaware of the danger. An observer stands at a switch that can direct the trolley down another track, but there is another

---

[113] Lin et al., *supra* note 98, at 27–41.

[114] WALLACH & ALLEN, *supra* note 59, at 16. Wallach also points out if robots are able to address ethical issues, new markets for robots will open up. In contrast, "if they fail to adequately accommodate human laws and values, there will be demands for regulations that limit their use." Wendell Wallach, *From Robots to Techno Sapiens: Ethics, Law and Public Policy in the Development of Robotics and Neurotechnologies*, 3 LAW INNOV. & TECH. 185, 196 (2011). *See also* Kenneth Kernaghan, *The Rights and Wrongs of Robotics: Ethics and Robots in Public Organizations*, 57 CANADIAN PUB. ADMIN. 485, 485 (2014) (arguing for the development of robots that follow ethical standards of personal moral responsibility, privacy, and accountability as robots become more commonplace in the areas of aging, public health, and defense).

[115] Lin et al., *supra* note 98, at 34.

[116] *Id.* at 41.

[117] WALLACH & ALLEN, *supra* note 59, at 117–24. The authors refer to studies that have discussed how neural network programming resonates with Aristotle's explanation of how people develop virtues. *Id.* at 121–23.

[118] Keith Abney, *Robotics, Ethical Theory, and Metaethics: A Guide for the Perplexed*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS, 35, 41–45 (Patrick Lin et al., eds., 2011).

person at the end of that section track. The observer has the choice of doing nothing and allowing five people to be killed or throwing the switch with the result that one person will die.[119] Research based on the Trolley Problem has led Joshua Greene to conclude that human beings tend to motivated by both utilitarian and deontological ethical impulses.[120] Most people will choose to turn the switch, but the answer will vary if changes are made to the scenario, for example, if the five workers are adults and the one person is a child.[121] For the computer programmer, one issue is whether autonomous machines can be programmed to engage in finely tuned moral reasoning, even if in the end, there will be no right answer to the dilemma.[122] The designer will be forced to resolve the dilemma one way or the other, and an issue is whether he or she can be held responsible for doing so.[123]

---

[119] *See* WALLACH & ALLEN, *supra* note 59, at 13–16; Nick Belay, *Robot Ethics and Self-Driving Cars: How Ethical Determinations in Software will Require a New Legal Framework*, 40 J. LEGAL. PROF. 119, 120 (2015).

[120] JOSHUA GREENE, MORAL TRIBES: EMOTION, REASON, AND THE GAP BETWEEN US AND THEM 113–28 (2013).

[121] Belay, *supra* note 119, at 120–21.

[122] In Wallach's view, designing machines to engage in moral decision making encompasses two problems:

> The first problem entails finding a computational method to implement norms, rules, principles or procedures for making moral judgements. The second is a group of related challenges that I refer to as frame problems. How does the system recognise that it is in an ethically significant situation? How does it discern essential from inessential information? How does the AMA estimate the sufficiency of initial information? What capabilities would an AMA require to make a valid judgement about a complex situation, eg, combatants vs. non-combatants? How would the system recognise that it had applied all necessary considerations to the challenge at hand or completed its determination of the appropriate action to take?

Wallach, *supra* note 114, at 200. For a discussion of how ethical control could be inserted into autonomous weapons systems, see Ronald C. Arkin, Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture, Georgia Institute of Technology Technical Report GIT-GVU-07-11 (2007), at 14–21, http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf.

[123] Belay, *supra* note 119, at 122–29. In an interesting study, Peter Danielson used a survey platform to seek responses to an autonomous machine version of the Trolley

Other literature is exploring ways in which machines can be made more amenable to punishment. As is true with other aspects of responsibility for autonomous machines, the question of punishment involves articulating the reasons for sanctions, whether new forms of punishment must be designed, and whether they would be technically feasible. J. Storrs Hall believes the deterrent function of punishment can be programmed into machines to influence behavior: "[i]n the rational machine . . . a credible threat of punishment (or reward) will be added to calculated utility of the predicted outcome of the act."[124] Asaro responds in part by arguing that Hall's deterrence approach fails to

---

Problem. Peter Danielson, *Surprising Judgments about Robot Drivers: Experiments on Raising Expectations and Blaming Humans*, 9 NORDIC J. APPLIED ETHICS 73 (2015). In this variation a train is being operated by a robot, which must decide between killing five people or turning to another track and killing one. Danielson acknowledges there are methodological issues with the study, but his results suggest people view the problem differently when a machine is involved. In another study, 90% of respondents answered that a human should divert the train. *Id.* at 78, *citing* John Mikhail, *Universal Moral Grammar: Theory, Evidence and the Future*, 11 COGNITIVE SCI. 143 (2007). However, in Daniel's study, only 37% agreed that the robot should divert the train, and more chose to be neutral rather than resolve the dilemma. Many expected the automated system to eliminate these kinds of problems. *Id.* at 78. More interesting, however, a surprising number of respondents blamed the victims in the problem for being on the track in the first place. *Id.* at 79. When a human was involved, no such blaming took place. *Id.* at 80. Finally, respondents were asked to respond to a situation when a child steps in front of a driverless car in a situation where it is physically impossible to stop the car. Most respondents found that the parents, the child, or the maker of the care should be held responsible. *Id.* This situation is an example of machines 'embodying' the ethical dilemmas with which humans wrestle and resolving them. The human response to machines resolving that dilemma is unexpected.

[124] J. Storrs Hall, Towards Machine Agency: A Philosophical and Technological Roadmap, "We Robot" Conference at the University of Miami Law School (Mar. 30, 2012), at 4, http://robots.law.miami.edu/wp-content/uploads/2012/01/Hall-MachineAgencyLong.pdf. Hall uses an example where a robot is in a situation in which there are two alternatives: to pick up a $5 bill or a $10 bill. Its utility function is the amount of money it has. It will pick up the $10. If we want the robot to pick up the $5 bill instead, the robot can be threatened with a $6 fine for picking up the $10 bill. The robot will then pick it up the $5 since it will net only $4 if it picks up the $10 bill. *Id.* Suppose the robot is given the choice between being placed in a situation where it can choose unencumbered or having its utility function changed so that the robot will prefer to pick up the $5 instead of $10. *Id.* It will choose the former because under its present utility function it will prefer to make $10. *Id.* Hall presumably uses the latter illustration to show that the threat of punishment in the form of changing a utility function can influence robot behavior.

encompass other reasons for punishment, such as retribution, deterrence that goes beyond the individual to the larger society, and reform.[125] Although Asaro himself does not suggest an alternative form of punishment, others have pointed to alternatives such as confining a robot to a particular part of cyberspace, deleting an autonomous machine's computer systems without backup, or banning a system from being used.[126] Sparrow, however, doubts that various forms of punishment will ever be completely satisfactory because in his view, punishment entails suffering.[127] This view raises another issue: whether it is ethical to design something that is capable of feeling something analogous to pain.

B. Autonomous Machines Having Legal Status or Personhood

The preceding subpart indicates some of the difficulties involved with programing ethical machines, which challenge current understandings of human consciousness, free will, etc. Law can avoid many of these issues since autonomous machines could be given legal status without answering these almost metaphysical questions. As discussed above, Chopra and White argue agency should be used to address harms caused by autonomous machines. With products liability, in their view, it would be useful to hold autonomous machines liable themselves, in part because they believe it will be difficult for plaintiffs to succeed under current products liability law.[128] Similarly, Pagallo suggests giving legal personhood to computer-based contracting systems to better justify holding contracts made by such systems enforceable.[129] The law could devise ways to create economic consequences for holding an autonomous machine responsible for harms, such as a minimum capital requirement associated with the

---

[125] Peter A. Asaro, *Punishment, Reinforcement Learning & Machine Agency*, COSMOPOLIS (Apr. 4, 2014), http://www.cosmopolis.globalist.it/Detail_News_Display?ID=69610.

[126] Pagallo, *supra* note 30 at 56 (confinement and deletion); Bernd Carsten Stahl, *Responsible Computers? A Case for Ascribing Quasi-Responsibility to Computers Independent of Personhood or Agency*, 8 ETHICS & INFO. TECH. 205, 211 (2006) (banning).

[127] Sparrow, *supra* note 27, at 72.

[128] CHOPRA & WHITE, *supra* note 10, at 143–144.

[129] PAGALLO, *supra* note 9, at 154.

machine; keeping a register that accounts for damage caused by a machine, presumably to be paid by its owner, operator, or some common fund; or using *respondeat superior* to hold a principal liable for what the machine has done.[130] Similarly, Lawrence Solum points out a computer agent could purchase insurance against the risk of its own misfeasance.[131]

Proponents of granting legal status or personhood to autonomous machines argue giving legal personhood to things is not new. Ships and corporations enjoy status as legal persons and assume liabilities. Chopra and White argue in several cases actions, not mental states, are important in determining legal liability. Further, they urge, mental states and intentionality are themselves constructions, which are used to determine when the same action is subject to legal sanctions and when it is not.[132] They also contend it is a categorical mistake to equate legal responsibility with moral responsibility—there is no need in their view to satisfy all the criteria for holding a robot morally responsible for something before it can be found legally responsible.[133]

Granting legal status to autonomous systems does have the advantages just discussed. At the same time, the mere grant of legal status does not resolve all moral issues. Some remarks by Mireille Hildebrant are interesting in this regard. Hildebrant agrees granting legal personhood to robots has less to do with recognizing something innate in the machine than with the consequences that follow from

---

[130] *Id.* at 103–106; CHOPRA & WHITE, *supra* note 10, at 150. It is unclear, however, whether such a response resolves the distributional problems posed by group responsibility. Unless autonomous machines are allowed to generate and retain their own income (a quasi-property right enjoyed by those machines), the funds used to satisfy third party claims would have to come from some human entity or group. Of course, if that group is a corporation, it could be argued that although shareholders would see less dividends, it is not inappropriate that ultimate responsibility would rest with them, since they benefit from the corporation's use of the machine.

[131] Solum, *supra* note 107, at 1245. He uses this experiment in response to a claim that an artificial intelligence could not be considered a legal person because it could not be held responsible in the sense of satisfying legal claims brought against it. *Id.* Solum acknowledges that insurance might not be available in all cases, leaving the artificial trustee unpunished. In his view, however, whether that disqualifies the trustee from being given legal personhood status depends in part on the purpose of punishment. *Id.* at 1245–47.

[132] CHOPRA & WHITE, *supra* note 10, at 146.

[133] *Id.* at 147.

granting such status.[134] She states that "moral agency is not necessarily the golden standard of personhood; if entities without such agency cause damage or harm it may be expedient and even justified to hold them accountable."[135] For Hildebrant, what warrants granting personhood is fairness to injured parties and to other perpetrators; "[t]he justification would reside in the ensuing obligation to compensate the damage or to contribute to the mitigation of the harm (justice done to the victim), but also *in the fairness of the distribution of liability* (justice in relation to other offenders)."[136] This argument implies that current forms of associative responsibility and their distribution among members of groups cannot adequately address situations when harm results from humans and machines working together, unless some kind of status is given to the machines themselves.

A combination of pragmatics and ethical quandaries about attributing responsibility by association could lead to the 'solution' of granting legal status to autonomous machines. The question then becomes whether such machines should be granted legal rights in addition to duties. The debate is remarkable in several respects. One reason is that it exists at all. Most of us do not think of machines as enjoying rights. Another reason is it highlights some of the conceptual difficulties that underlie them. On the one hand, conferring rights on machines would seem to confirm a positivistic, constructive understanding of rights. This understanding of course raises the issue whether such rights can then be taken away. On the other hand, proponents of inherent rights could argue truly autonomous machines must possess some quality, such as intelligence, that it shares with humans who enjoy rights. However, this argument raises the problem of essentialism: identifying what makes human beings rights-bearing persons. Assuming this quality can be identified, it raises the issue

---

[134] Mireille Hildebrandt, *From Galatea 2.2 to Watson—And Back?* in HUMAN LAW AND COMPUTER LAW: COMPARATIVE PERSPECTIVES, 25 JUS GENTIUM 23, 38 (Mireille Hildebrandt & Jeanne Gaakeer, eds., 2013). In this regard, Bernd Stahl uses the distinctions between various forms of responsibility identified by scholars such as Watson and Smith discussed in Part II.C to argue robots can be assigned responsibility to further certain social ends. Bernd Carsten Stahl, *Responsible Computers? A Case for Ascribing Quasi-Responsibility to Computers Independent of Personhood or Agency*, 8 ETHICS & INFO. TECH. 205, 210–11 (2006).

[135] Hildebrant, *supra* note 134, at 38.

[136] *Id*. (emphasis added).

whether rights should be given to all entities that share it.[137] A third issue is how older concepts of hierarchy and status are being used to frame the problem. It could be argued intelligent machines should not be given rights because machines are like animals or children, beings with diminished capacities who do not enjoy the full rights of adult humans for that reason.[138] Commentators have also used slave law to discuss how human 'masters' might be held responsible for the actions of robot 'slaves'[139] and to discuss why autonomous machines should not enjoy the same rights as humans.[140]

The moral concerns raised by this third feature of the debate over rights for machines, combined with the tendency of humans to anthropomorphize machines[141] and efforts to program computers to be emotionally intelligent in their interactions with human beings,[142] might finally lead to arguments that the most highly sophisticated machines are owed moral consideration. For Sparrow this is a very high bar. He posits a moral "Turing test" such that robots would merit full moral status only when robots display properties so that it would be difficult to choose between the life of a human being and the existence of the autonomous machine.[143] David Gunkel, however, evaluates other ways

---

[137] For a discussion of these problems, see Solum, *supra* note 107 at 1262-74.
[138] Asaro, *A Body to Kick*, *supra* note 16, at 178.
[139] See CHOPRA & WHITE, *supra* note 10, at 134; Lin et al., *supra* note 98, at 66.
[140] Asaro, *A Body to Kick*, *supra* note 16, at 178; Solum, *supra* note 107, at 1279 (criticizing the slave argument).
[141] Mark Coeckelbergh argues in this regard that we anthropomorphize robots as part of a hermeneutic through which we view robots as individuals. Mark Coeckelbergh, *Is Ethics of Robots About Robots? Philosophy of Robotics Beyond Realism and Individualism*, 3 LAW INNOVATION & TECH. 241, 244 (2011).
[142] *See* Brian R. Duffy, *Anthropomorphism and the Social Robot*, 42 ROBOTICS & AUTONOMOUS SYS. 177 (2003) (arguing the tendency to anthropomorphize will actually assist in developing machines that enhance meaningful interactions with humans). For an example of how some robots are designed to elicit emotional responses in humans, see Eun Ho Kim et al., *Design and Development of an Emotional Interaction Robot, Mung*, 23 ADVANCED ROBOTICS 767 (2009) (describing a robot designed to emulate bruising); Hawon Lee & Eunja Hyun, *The Intelligent Robot Contents for Children with Speech-Language Disorder*, 18 EDUC. TECH. & SOC'Y 100 (2015) (describing a robot used to work with children with speech and language disabilities).
[143] Robert Sparrow, *The Turing Triage Test*, 6 ETHICS INFO. TECH. 203 (2004); Robert Sparrow, *Can Machines Be People? Reflections on the Turing Triage Test*, *in* ROBOT

of framing the issue. He describes work drawn from the animal rights and environmental rights movements to explore whether machines might be entitled to moral patiency. This approach is not concerned with whether someone or something has moral agency with rights and responsibilities, but asks instead whether that entity can suffer.[144] Gunkel cites the work of Lucian Floridi, who argues that information is the common denominator between animals, the environment, and computers: "[w]hat makes someone or something else a moral patient, deserving of some level of ethical consideration (no matter how minimal), is that it exists as a coherent body of information."[145] Under this view, the loss of information, a form of informational entropy, is analogous to suffering.[146] Gunkel himself is critical of this approach because in his view, finding a common denominator between humans and other entities to justify giving moral regard to nonhumans is a form of essentializing that does violence to differences in the individuals that are part of the group and excludes others.[147] He therefore explores an ethic based on concern for the other drawn from a variety of scholars influenced by Levinas[148] so that the machine is included among those others that demand moral consideration. However, for Gunkel, this approach is also flawed because it falls back on a kind of exclusion in which machines are always left out.[149] For Gunkel, the question whether machines should be given moral consideration is not one that can ultimately be answered, rather one that should constantly be asked because of the light it sheds on our conceptions of ethics.[150]

V. CONCLUSION

This Article began with Asaro's call for theories of responsibility that can address large, complex systems of human beings and machines working together, so those systems will yield desirable

---

ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS, 301, 301 (Patrick Lin et. al. eds., 2012).

[144] DAVID J. GUNKEL, THE MACHINE QUESTION: CRITICAL PERSPECTIVES ON AI, ROBOTS, AND ETHICS 93–157 (2012).

[145] *Id.* at 146.

[146] *Id.*

[147] *Id.* at 157.

[148] *Id.* at 177.

[149] *Id.* at 206–07.

[150] *Id.* at 211.

outcomes and can be held responsible when the results are otherwise. I have responded by setting out a possible trajectory for the co-evolution of legal responsibility and autonomous machines. Commentators are using pre-existing legal doctrines related to defective products, agency law, and international humanitarian law. They debate the extent to which those doctrines in their current forms can address situations that will arise when autonomous machines become more common. If existing law does not provide satisfactory solutions, it is because of the law's general discomfort with associative responsibility, a discomfort that is shared and supported by most of the literature on ethics. The ethical literature most relevant to the problems of associative responsibility provides some guidance on the issue, but no completely satisfactory answers. In turn, the concern there will be gaps in responsibility for harms caused by machines leads to two lines of development. The first is refining or redefining the concept of responsibility. The second is to reduce harm by designing autonomous machines with prosocial behaviors. If successful, that very success, combined with calls to grant legal personhood to machines for legal and pragmatic reasons and the human tendency to anthropomorphize, will strengthen what are now nascent calls to treat such machines as moral agents.

To what extent does tracing a possible co-evolutionary trajectory respond to Asaro's call? To answer that question, it is helpful to toggle back and forth between various points in that trajectory. At the end point, a world in which humans and machines who enjoy equal legal status and rights would of course be radically different: for the first time in human history, we would co-exist with nonhuman intelligences who are our equals (and perhaps our superiors) in significant ways. We will have created our own alien 'life.' However, there is a sense in which we could use our current systems of legal responsibility without much controversy: the autonomous machine would be treated like any other individual who lives and works in large systems.

Asaro's challenge can thus be reframed: how well do our legal doctrines address harms caused by complex systems of humans now, with or without machines? I have discussed that latter question to some extent in Part II. A complete answer to that question might be a matter of the glass being half empty or half full. The debates surrounding tort reform serve as an example. In a 1994 meta-study, Gary Schwartz surveyed then-existing assessments of the impact of tort law in a wide area of economic sectors and concluded there was "evidence

persuasively showing that tort law achieves something significant in encouraging safety."[151] However, the impact of tort law can sometimes be ambiguous or of lesser importance than other factors. A study by Paul Rubin, for example, indicates that consumer preferences for safer products are the primary drivers of improvements in safety. In his view, regulation and tort law can also contribute to safety improvements. However, because tort law is an expensive means of encouraging safety, it might actually increase risk by causing people to forgo things such as drugs and medical treatments because they are made more expensive by costs incurred to avoid tort liability.[152]

It is beyond the scope of this Article to assess the effectiveness of tort law, but for our purposes, it is enough to repeat what was discussed earlier: the law already purports to address large systems. Assessing how effective the law is in regulating those systems does not require us to take autonomous machines into account. However, this is not to say autonomous machines are irrelevant. The extent to which they do become relevant will depend on how deeply societies want to penetrate complex systems to hold parts of those systems responsible for harms. In this regard, Wallach points out that the investigation of the *Challenger* disaster demonstrates how hard it is to determine who or what is to blame for the failure of a complex system such as the space shuttle, that in turn is a product of complex organizations like large corporations. [153] It is only if society feels it is necessary to become finer grained in assigning responsibility to move from the corporations who manufactured and designed the components and software used in the shuttle to individual designers and engineers who could be said to have contributed to the defects that led to that disaster, as well those along the chain of command that ordered the launch to go forward, that the problems of associational responsibility discussed in this Article become more salient. Autonomous machines then would become part of the calculus, if by that time their decision-making capacity is so sophisticated that it will be hard to attribute responsibility for harms they cause to their coworkers, supervisors, or those who designed them,

---

[151] Gary T. Schwartz, *Reality in the Economic Analysis of Tort Law: Does Tort Law Really Deter?*, 42 UCLA L. REV. 377, 423 (1994–1995).
[152] Paul H. Rubin, *Markets, Tort Law, and Regulation to Achieve Safety*, 31 CATO J. 217, 231–32 (2011). Rubin argues that tort reform from 1981 to 2000 led to 24,000 fewer accidental deaths because of increased emergency medical care. *Id.*
[153] Wallach, *supra* note 114, at 194–95.

but at the same time, they are not autonomous enough to merit legal, let alone moral, agency so that they can be blamed directly for what they have done.

At that point, prevailing ethical and legal views of responsibility would need to be reevaluated. Part III has assessed possible ways existing conceptions of responsibility might be modified to encompass humans and machines. In my view, however, those modifications require the creation of new or more abstract ethical and legal subjects capable of bearing responsibility that will involve necessarily some form of responsibility by association. Many of us will find that hard to accept, although some change might be possible at the margins. This view means the strategy of designing machines themselves with a view towards harm reduction will be seen as more desirable, with the possible implications discussed in Part IV. One of the ironies of that approach, however, is that our aversion to sharing the responsibility of another could lead to the development of machines that are in some senses wholly other and in other senses wholly us.