8

HOW TO PUNISH A ROBOT[1]AND[2]

Mark A. Lemley[3]
Bryan Casey[4]

What happens when artificially intelligent robots misbehave? The question is not just hypothetical. As robotics and artificial intelligence systems increasingly integrate into our society, they will do bad things. In this Essay, we explore some of the challenges emerging robotics technologies will pose for remedies law. We argue robots will require us to rethink many of our current doctrines and that the emerging technology also offers important insights into the law of remedies we already apply to people and corporations.

[2] **Como citar este artigo científico.** LEMLEY, Mark A.; CASEY, Bryan. How to punish a robot. In: **Revista Amagis Jurídica**, Ed. Associação dos Magistrados Mineiros, Belo Horizonte, v. 13, n. 3, p. 195-229, set.-dez. 2021.

[3] William H. Neukom Professor, Stanford Law School; partner, Durie Tangri LLP. William H. Neukom, Stanford Law School. Director, Stanford Program in Law, Science, and Technology. Senior Fellow, Stanford Institute for Economic Policy Research. Affiliated Professor, Stanford Symbolic Systems Program. California, United States. *e-mail*: mlemley@law.stanford.edu or mlemley@durietangri.com

[4] Research Fellow, Center for Automotive Research at Stanford (CARS).

INTRODUCTION

Engineers training an artificially intelligent self-flying drone were perplexed. They were trying to get the drone to stay within a predefined circle and to head toward its center. Things were going well for a while. The drone received positive reinforcement for its successful flights, and it was improving its ability to navigate toward the middle quickly and accurately. Then, suddenly, things changed. When the drone neared the edge of the circle, it would inexplicably turn *away* from the center, leaving the circle.

What went wrong? After a long time spent puzzling over the problem, the designers realized that whenever the drone left the circle during tests, they had turned it off. Someone would then pick it up and carry it back into the circle to start again. From this pattern, the drone's algorithm had learned – correctly – that when it was sufficiently far from the center, the optimal way to get back to the middle was to simply leave it altogether. As far as the drone was concerned, it had discovered a wormhole. Somehow, flying outside of the circle could be relied upon to magically teleport it closer to the center. And far from violating the rules instilled in it by its engineers, the drone had actually followed them to a T. In doing so, however, it had discovered an unforeseen shortcut – one that subverted its designers' true intent.

What happens when artificially intelligent robots don't do what

we expect, as the drone did here? The question is not just hypothetical. As robotics and artificial intelligence (AI) systems increasingly integrate into our society, they will do bad things. Sometimes they will cause harm because of a design or implementation defect: we should have programmed the self-driving car to recognize a graffiti--covered stop sign but failed to do so. Sometimes they will cause harm because it is an unavoidable by-product of the intended operation of the machine. Cars, for example, kill thousands of people every year, sometimes unavoidably. Self-driving cars will too. Sometimes the accident will be caused by an internal logic all of its own – one that we can understand but that still doesn't sit well with us. Sometimes robots will do the things we ask them to (minimize recidivism, for instance) but in ways we don't like (such as racial profiling). And sometimes, as with our drone, robots will do unexpected things for reasons that doubtless have their own logic, but which we either can't understand or predict.

These new technologies present a number of interesting questions of substantive law, from predictability, to transparency, to liability for high-stakes decision-making in complex computational systems. A growing body of scholarship is beginning to address these types of questions. Our focus here is different. We seek to explore what remedies the law can and should provide once a robot has caused harm.

The law of remedies is transsubstantive. Whereas substantive law defines who wins legal disputes, remedies law asks, "What do I get when I win?" Remedies are sometimes designed to make plaintiffs whole by restoring them to the condition they would have been in "but for" the wrong. But they can also contain elements of moral judgment, punishment, and deterrence. For instance, the law will often act to deprive a defendant of its gains, even if the result is a windfall to the plaintiff, because we think it is unfair to let defendants keep those gains. In other instances, the law may order defendants to

do (or stop doing) something unlawful or harmful.

Each of these goals of remedies law, however, runs into difficulties when the bad actor in question is neither a person nor a corporation but a robot. We might order a robot – or, more realistically, the designer or owner of the robot – to pay for the damages it causes. (Though, as we will see, even that presents some surprisingly thorny problems.) But it turns out to be much harder for a judge to "order" a robot, rather than a human, to engage in or refrain from certain conduct. Robots can't directly obey court orders not written in computer code. And bridging the translation gap between natural language and code is often harder than we might expect. This is particularly true of modern AI techniques that empower machines to learn and modify their decision-making over time, as the drone in the opening example did. If we don't know how the robot "thinks," we won't know how to tell it to behave in a way likely to cause it to do what we actually want it to do.

One way to avoid these problems may be to move responsibility up the chain of command from a robot to its human or corporate masters – either the designers of the system or the owners who deploy it. But that too is easier said than done. Robot decision-making is increasingly likely to be based on algorithms of staggering complexity and obscurity. The developers – and certainly the users – of those algorithms won't necessarily be able to deterministically control the outputs of their robots. To complicate matters further, some systems – including many self-driving cars – distribute responsibility for their robots between both designers and downstream operators. For systems of this kind, it has already proven extremely difficult to allocate responsibility when accidents inevitably occur.

Moreover, if the ultimate goal of a legal remedy is to encourage good behavior or discourage bad behavior, punishing owners or designers for the behavior of their robots may not always

make sense – if only for the simple reason that their owners didn't act tortiously. The same problem affects injunctive relief. Courts are used to ordering people and companies to do (or stop doing) certain things, with a penalty of contempt of court for noncompliance. But ordering a robot to abstain from certain behavior won't be trivial in many cases. And ordering it to take affirmative acts may prove even more problematic. In this Essay, we outline the basic principles of remedies law and consider how remedies will – or won't – work when applied to robots.

I REMEDIES AND ROBOTS

I.A. THE NATURE OF REMEDIES

A.1 NORMATIVE VERSUS ECONOMIC PERSPECTIVES

The choice of remedy for a given legal violation often stems from fundamental assumptions regarding the nature of the substantive law itself. Two views predominate. A "normative" view of substantive law sees it as a prohibition against certain conduct, with the remedy being whatever is prescribed by the law itself. The defendant, on this view, has engaged in a wrongful act that we would stop if we could. But because it is not always possible to do so – commonly because the act has already occurred – remedies law seeks to do the next best thing: compensate the plaintiff for the damage done.

An alternative view of substantive law, however, conceptualizes the role of remedies differently. Under this "economic" view, the substantive law alone forbids nothing. Rather, it merely specifies the foreseeable consequences of various choices, with the available remedies signaling the particular penalties associated with

particular conduct. Damages, on this view, are simply a cost of doing business – one we want defendants to internalize but not necessarily to avoid the conduct altogether.

To help illustrate the difference, consider an everyday encounter with a traffic light. Under the normative view, a red light stands as a prohibition against traveling through an intersection, with the remedy being a ticket or fine against those who are caught breaking the prohibition. We would stop you from running the red light if we could. But because policing every intersection in the country would be impossible, we instead punish those we do catch in hopes of deterring others.

Under the economic view, however, an absolute prohibition against running red lights was never the intention. Rather, the red light merely signals a consequence for those who do, in fact, choose to travel through the intersection. As in the first instance, the remedy available is a fine or a ticket. But under this view, the choice of whether or not to violate the law depends on the willingness of the lawbreaker to accept the penalty.

In one of his more arresting turns of phrase, Justice Oliver Wendell Holmes Jr famously described the economic view of substantive law as that of a "bad man." According to Justice Holmes:

> If you want to know the law and nothing else, you must look at it as a bad man, who cares only for the material consequences which such knowledge enables him to predict, not as a good one, who finds his reasons for conduct, whether inside the law or outside of it, in the vaguer sanctions of conscience (HOLMES JR, 1897, p. 457-459).

The measure of the substantive law, in other words, is not to be mixed up with moral qualms, but is simply coextensive with its remedy – no more and no less.

While some law and economics scholars accept this precept as fundamental, in many behavioral contexts it does not tell the entire story. Although the actual consequences associated with lawbreaking play a substantial role in much of human decision-making, many individuals nonetheless view law as having distinctly normative underpinnings. As Laycock notes, "It is certainly true that some individuals will obey the law only if the consequences of violation are more painful than obedience," but the fact that "some individuals are unmoved does not eliminate the statement's moral force for the rest of us." (LAYCOCK, 2020, p. 3-7).

A.2 BAD MEN AND GOOD ROBOTS

People are rarely forced to grapple with the distinctions between the normative or economic view of substantive law. But robots, or their programmers at least, are afforded no such luxury. Sure, robots can be prohibited from engaging in certain types of conduct, assuming their designers understand and control the algorithm by which they make decisions. But implementing a legal remedy via computer code necessarily involves adopting either a normative or economic view of the substantive law.

That's because a true "prohibition" can only be communicated to a computer system in one of two basic ways: it can be encoded in the form of an "IF, THEN" statement that prevents a robot from engaging in particular types of conduct, or it can be coded as a negative weight for engaging in that same conduct. An IF, THEN statement operates like an injunction, while a weight in a decision-making algorithm operates like a liability rule.

Returning to the example of the red light, a programmer seeking to prohibit a robot from breaking the law could do so with an IF, THEN statement along the lines of: "If the robot

encounters a red light, then it will not travel into the intersection." Similarly, a programmer seeking to achieve that same prohibition in a probabilistic system could do so by assigning an infinitely high negative consequence to traveling into the intersection when the light is red.

An IF, THEN statement is an absolute rule. If a triggering event occurs, then a particular consequence must inexorably follow. As a practical matter, so is an infinitely negative weight. Both achieve the functionally equivalent result of prohibiting the unlawful conduct – the goal of a normative vision of substantive law. But in order to achieve this normative vision, the prohibition must be implemented without regard for the cost of a ticket.

Because the law is encoded as an absolute in its programming, the robot will always obey the law. That's not true of people. If we want legal rules to be self-executing, the ability to impose perfect obedience may be a good thing.

By contrast, if the underlying theory of a remedy is economic, the machine's decision-making calculus is fundamentally different. Once more, the example of the traffic light helps to clarify this distinction. To an economist, the substantive law and its remedy do not signal a "self-executing refusal to ever run a red light" but instead an understanding that "running a red light is associated with a small chance of a modest fine and a somewhat increased chance of a traffic accident which will damage the car and may require the payment of damages to another." Under this view, the remedy, and its risks, are both expressed in probabilistic terms. They translate into probabilistic costs within the robot's overall decision-making calculus. Those costs won't be infinite, unless perhaps the penalty is death. They will instead reflect a "price" for running a red light that the algorithm might decide to pay depending on what benefits light--running offers.

Thus, under the economic view, the choice of whether to obey a law is, of necessity, the choice of a Holmesian "bad man." Normative views of substantive law – which we know shape certain aspects of human behavior – cannot be expected to translate cleanly into the robotics context with their associated remedies intact. If we want robots to adopt normative views of the law, we will need outright prohibitions. And imposing prohibitions rather than costs will make it hard for robots to achieve many things. After all, it's hard to operate a robot with too many absolute prohibitions.[5] And this will be particularly true of machine learning systems that develop their own algorithms, making it difficult for engineers to reliably predict how encoded prohibitions will interact with other rules.

Encoding the rule "don't run a red light" as an absolute prohibition, for example, might sometimes conflict with the more compelling goal of "not letting your driver die by being hit by an oncoming truck." Humans know that "don't run a red light" doesn't really mean "don't *ever* run a red light." Rather it translates, roughly, to "don't run a red light unless you have a sufficiently good reason and it seems safe." Likewise, even weightier normative prohibitions, such as "thou shalt not kill," come with an implied "unless. [...]" But designers can't put that in an IF, THEN statement unless they understand and specify all the exceptions to the rule.

More plausibly, robots operating in the real world will have to adopt algorithmic approaches to almost all complex problems that weigh particular actions against various goals and risks. As a result, the role of remedies in discouraging socially detrimental conduct will need to be reimagined in terms of cost internalization,[6] as opposed to normative sanction or punishment. Deterrence makes

[5] "Don't become Skynet" does seem like a good one to include, though.

[6] By "internalization," we do not necessarily mean that the law should attempt to put an explicit monetary value on every conceivable form of harmful conduct. Rather, internalities and externalities can be addressed by a multitude of direct *and* indirect means, just as the law does today.

sense where we are trying to affect individual behavior. But the logical way to "deter" a machine is to put the actual costs into the calculus it uses to make the decision. In practice, that translates into quantifying, and then operationalizing, the price we want robots to have to pay if they take certain actions we want to deter. And under the broadest interpretation of the economic view, even doctrines seemingly designed to prevent or deter conduct – like injunctions or prison sentences – could simply be construed as costs, albeit very high ones.

That said, we think it makes more sense to distinguish between remedies designed to internalize costs and those designed to enjoin, deter, or punish behavior. While some defendants faced with the latter may treat punitive damages or even prison sentences as mere costs of doing business, the remedy's ultimate intent is to deter unlawful conduct, not to simply internalize its social costs.

For the vast majority of applications, legal remedies will likely be incorporated into machines through their "economic" formulation – resulting in robots that, by design, adopt this view of substantive law exclusively. Unless specifically programmed otherwise, distinctions between normative and economic goals will be utterly lost on robots. Thus, while it may be true to say that it is the rare "individual [...] [who] will obey the law only if the consequences of violation are more painful than obedience," (LAYCOCK, 2020, 7). this will be definitionally true of robots. And for reasons made clear in virtually every sci-fi plot line featuring robots, it will only be on the rarest of occasions that it actually makes sense to completely bar robots from engaging in certain types of conduct.

It thus appears that Justice Holmes's archetypical "bad man" will finally be brought to corporeal form, though ironically, not as a man at all. And if Justice Holmes's metaphorical subject is truly "morally impoverished and analytically deficient," as some accuse, it

will have significant ramifications for robots (consider BEZEMEK, 2016, p. 15).

B TEACHING ROBOTS TO BEHAVE

Each of the major types and purposes of remedies identified above will face challenges as applied to robots and AI. We consider each in turn below.

B.1 WHO PAYS?

The first purpose of damages – to compensate plaintiffs for their losses and so return them to their rightful position – is perhaps the easiest to apply to robots. True, robots don't have any money. So they generally can't actually pay damage awards themselves.

But this problem is hardly insurmountable. The law will rise to the challenge. Someone built the robots, after all. And someone owns them. So if a robot causes harm, it may make sense for the company behind it to pay, just as when a defective machine causes harm today.

But it's not that easy. Robots are composed of many complex components, learning from their interactions with thousands, millions, or even billions of data points, and they are often designed, operated, leased, or owned by different companies. Which party is to internalize these costs? The one that designed the robot or AI in the first place? The one that collected and curated the data set used to train its algorithm in unpredictable ways? The users who bought the robot and deployed it in the field? Sometimes all of these roles will be one in the same, falling upon individuals operating in a single company.

Robot designers, owners, operators, and users will, of course, fight over who bears true legal responsibility for causing the robot

to behave the way it did. And these complex distinctions don't even account for the role of third parties causing robots to behave in adverse ways, as recently happened when Microsoft's chatbot, Tay, turned into a proverbial Nazi after interacting with trolls on Twitter (see VINCENT, 2016).

These problems aren't new, of course. Suppliers in a product chain have blamed each other when things go wrong for a long time, and courts have had to sort those claims out. Responsibility issues for robots too can, and will, eventually be resolved by the courts. But long before any consensus is reached, we should expect no shortage of finger-pointing, as different companies and individuals clamor to shift responsibility for harms to others in the causal chain – whether just to minimize their costs or because there are legitimate disputes about how the behavior of different actors in the chain interacted to cause the harm.

B.2 LAW AS ACTION: SHAPING THE BEHAVIOR OF *RABOTA ECONOMICUS*

The second prong of the remedies triad – damage awards and equitable remedies designed to internalize costs and deter socially unproductive behavior – will likely prove even more problematic. If we want to deter a robot, we need to make sure that it is programmed to account for the consequences of its actions. Embedding this type of decision-making in robots often means quantifying the various consequences of actions and instructing the robot to maximize the expected net monetary benefits of its behavior.

This might sound like heaven to an economist. Finally, we will have a truly rational *homo economicus* (or, more accurately, a *rabota economicus*) (see SCIENCE, 2011) who will internalize the social costs of its actions (at least insofar as those costs are accurately calculated) and modify its behavior accordingly. And if

machine learning systems estimate these costs correctly, robots will be "Learned" indeed – presumably deciding to do harm only when it is socially optimal (that is, when B < PL).[7]

But not so fast. Things are more complicated. Robots won't reflexively care about money. They will do whatever we program them to do. We can align robot incentives with social incentives by properly pricing, punishing, or deterring the companies that design, train, own, or operate robots. Those companies, in turn, should internalize the relevant costs of their robots' actions. It might be reasonable to assume that corporations and people want to maximize their rational self-interest and will, thus, program their robots accordingly. But not all will, either intentionally or unintentionally. There are at least three potential problems.

First, the goal of cost internalization through legal liability can only be accomplished by proxy. And it isn't clear who the proxy will be. All the problems we noted in the prior section about assigning responsibility to compensate victims will return in spades as we try to force robots to account for the costs of their conduct. Even truly rational, profit-maximizing companies with perfect information about the costs of their actions won't internalize those costs unless they expect the legal system to hold them liable. If they are wrong, either in fearing liability when none exists or in believing someone else will foot the bill, their pricing will not accurately reflect reality.

Second, we are unlikely to have anything resembling "perfect" information about the potential harms robots may cause. Getting robots to make socially beneficial, or morally "right," decisions means we first need a good sense of all the things that could go wrong. Unfortunately, we're already imperfect at that. Then we'd need to decide whether the conduct is something we want to ban, discourage, tax, or simply permit. Having done so, we would then

[7] See "United States v Carroll Towing Co", 159 F2d 169, 171–73 (2d Cir 1947).

need to decide who in the chain of robot design, training, ownership, and operation should be responsible for the harm, if anyone. Then, we would need to figure out how likely each adverse outcome is in any given situation. Finally, we would need to assign a price to those potential harms – even the amorphous ones, such as a reduction in consumer privacy. And we'd want to balance those harms against reasonable alternatives to make sure the decision the robot made was the right one, even if it did cause harm.

Our entire system of tort law has been trying to accomplish this feat for centuries. And it hasn't worked very well. Indeed, most of tort is composed of standards, as opposed to hard and fast rules, for good reason. Standards give us the leeway to reserve judgment for later, when we might have a better idea of the actual facts leading up to an event. Tort law, for example, requires us to value injury, and – if we are to deter conduct – to decide on a multiplier to that value that serves as an optimal deterrent. While there are some circumstances in which we calculate these values formulaically (see, for example, ROSS, 1980, 133-135), the primary way we do so is by leaving it to juries to pick the right number after an injury has already occurred. And we know virtually nothing of how juries will react to harmful events caused by robots.

The problem is even more complex than that, though, because robots don't necessarily care about money. They will maximize whatever they are programmed to maximize. If we want them to internalize the costs of their behavior, we will need to put those costs in terms robots can understand – for example, as weights that go into a decision-making algorithm. That's all well and good for robots already designed to maximize profit in purely monetary terms – say, a day-trading AI. But lots of robots will be designed with something other than money in mind. A policing or parole algorithm might minimize the likelihood that a released offender commits another crime. A self-driving car might minimize time to destination subject

to various constraints like generally obeying traffic laws and reducing the risk of accidents. But to build deterrence into those algorithms, we must convert certain divergent values into a common metric, whether it be money or something else.

A third complexity involving *rabota economicus* emerges for economic costs that are not directly reflected by legal remedies. The cost of any given decision, after all, is not just a function of the legal system. In many instances, extralegal forces such as ethical consumerism, corporate social responsibility, perception bias, and reputational costs will provide powerful checks on profit-maximizing behaviors that might, otherwise, be expected to produce negative societal externalities. By pricing socially unacceptable behavior through the threat of public backlash, these and other market forces help to fill some of the gaps left by existing remedies regimes. But they may open up other holes, creating rather than internalizing externalities.

Corporations are also likely to be siloed in ways that interfere with effective cost internalization. Machine learning is a specialized programming skill, and programmers aren't economists.[8] Even those who are employed by profit-maximizing companies interested in effectively internalizing their legal costs may see no reason to take the law into account, or may not be very good at it even if they try to. They may resent constant interference from the legal department in their design decisions. And agency costs mean that different subgroups within companies may be motivated by different incentives – as when sales divisions, manufacturing divisions, and service departments all get compensated based on different and potentially conflicting metrics.

Engineers aren't the only people whose motivations we need to worry about. What a self-learning robot will maximize depends

[8] At least not most of them.

not only on what it is designed to do – the default optimizing function or functions that it starts with – but also how it learns. To efficiently deter behavior, we must be able to predict it. But if we don't know how the robot will behave because it might discover novel ways of achieving the goals we specify, simply pricing in the cost of bad outcomes might have unpredictable effects, such as shutting down a new and better approach that produces some bad results but is nonetheless worth it. This complex relationship between deterrence, responsibility, and financial liability does not, alone, differentiate robots from corporations or people. Deterrence is imperfect among humans, too, because humans aren't motivated entirely by money and because they can't always pay for the harm they cause. But what is different here is that the possibility of deterrence working *at all* will depend entirely on the robot's code. A robot programmed to be indifferent to money won't be deterred by any level of legal sanction. And while making the responsible legal party pay[9] might encourage that party to design robots that do take adequate care, the division of responsibility between component makers, software designers, manufacturers, users, owners, and third parties means that the law must be careful about who exactly it holds accountable.

C DETERRENCE WITHOUT RATIONAL ACTORS: IS THERE STILL A ROLE FOR MORALITY AND SOCIAL OPPROBRIUM IN ROBOT REMEDIES?

C.1 EQUITABLE MONETARY RELIEF AND PUNISHMENT

So far, we have focused on internalizing the costs of accidents or other injuries that result from otherwise socially desirable activities, such as driving cars. But we also need to worry about genuinely "bad" behavior by robots that may merit prohibition. Many of our

[9] Or face time behind bars.

equitable monetary remedies are aimed at this sort of conduct. Their goal is not to make defendants internalize costs – to put a price on socially valuable behavior because of the costs it imposes – but to prevent the behavior. If you steal my car, the law says that you don't get to keep it even if you value it more than me. Rather, you hold it in constructive trust for me. If you make profits by infringing my copyright or trade secret (but not my patent), the law will require you to disgorge those profits, paying me the money you made even if I never would have made it myself. We require defendants to give up such "unjust enrichment," not because we think we need to do so to compensate the plaintiff, but because we don't want the defendant to have the money.

These equitable rules share some similarities with the cost-internalization measures discussed in the last Section. But there are two key differences: (1) the money a defendant must pay is not limited to what is needed to compensate the plaintiff, and (2) the defendant must give up all gains, making the entire activity unprofitable. The focus here is not on the plaintiff's rightful position but on the defendant's rightful position. And in the class of cases in which we often use these remedies, the defendant's rightful position is one in which she didn't engage in the activity at all.

From an economic perspective, depriving defendants of their gains is simply a matter of coming up with a number. It might be greater than, equal to, or less than the damages we would otherwise impose to internalize the costs of unlawful conduct or to restore the plaintiff's rightful position. But there is something psychologically effective about taking away a defendant's gains altogether. Indeed, in certain contexts, it might be a better means of deterring humans than the threat of paying compensatory damages, even if those damages turn out to be higher than a disgorgement remedy would. When it comes to robots, however, there is little reason to think that the notion of taking "all your profits" will have the same psychological

effects. True, if you set "profit = 0," a profit-maximizing AI would not engage in the conduct. But that same logic would apply with equal force if the damages award made the activity unprofitable too.

Remedies focused on the defendant's rightful position do have one significant economic advantage over damages remedies intended strictly as ex ante deterrents: we can calculate them after the fact once we have all the necessary information. If we want to use the threat of damages to deter conduct, we need to predict the likelihood and severity of the harm that the conduct will cause. But if we care only about depriving the defendant of benefits on the theory that doing so will deter her, we just need to wait to set the number until the parties get to court and figure out how much the defendant actually gained. That often won't be trivial. The benefit of stealing a trade secret, for example, can be as amorphous as a "quicker time to market" or a "more competitive product." (LEMLEY, 2017, p. 266-269). But it's still likely to be easier than predicting in advance who will be injured and by how much.

This same calculus doesn't work for injuries that are the by-product of productive behavior. It doesn't make sense to say that a self-driving car that hits a pedestrian should disgorge its profits. It likely didn't profit from hitting the pedestrian. And we don't want to force defendants to disgorge all the value they make from driving. But defendant-focused equitable monetary remedies, like disgorgement or constructive trust, may have advantages for robot torts for which our goal is to stop the conduct altogether, not simply to price it efficiently.

C.2 DETECTION, DETERRENCE, AND PUNITIVE DAMAGES

The fact that robots won't be affected by the psychological impact of certain remedies also has consequences for how we should

think about the threat of detection. For a robot to be optimally deterred by remedies like disgorgement – which rely on human psychology to maximize their effects – we must also detect and sanction the misconduct 100 percent of the time. That, in turn, leads us to the problem of robots (or their masters) that hide misconduct.

To be sure, many robot harms will be well-publicized. The spate of autonomous vehicle accidents covered by media in recent years provides one stark example. But countless robot harms will be of far subtler, so-called black box, varieties and will, therefore, be much harder to detect.

Makers and trainers of robots may have incentives to hide their behavior, particularly when it is profitable but illegal. If a company's parole algorithm concludes (whether on the merits of the data or not) that black people should be denied parole more often than similarly situated white people, it might not want the world to know. And if you, as an owner, tweaked the algorithm on your car to run over pedestrians rather than put your own life at risk, you might seek to hide that too. We have already seen remarkable efforts by companies conspiring to cover up wrongdoing, many of which succeeded for years. Often such conspiracies are brought down by sheer virtue of their scale – that is, the fact that many people know about and participate in the wrongdoing. This same property may be less true of future robotics firms, which may require fewer people to participate and cover up unlawful acts.

Further, robots that teach themselves certain behaviors might not know they are doing anything wrong. And if their algorithms are sophisticated enough, neither may anyone else for that matter. Deterrence will work on a robot only if the cost of the legal penalty is encoded in the algorithm. A robot that doesn't know it will be required to disgorge its profits from certain types of conduct will not accurately price those costs and so will optimize for the wrong behaviors.

The economic theory of deterrence responds to the improbability of getting caught by ratcheting up the sanctions when you are caught, setting the probability of detection times the penalty imposed equal to the harms actually caused. Proportionality of punishment makes sense here. As the chance of detection goes down we want the damage award to go up. And machines can do this math far better than humans can. Indeed, this idea may be tailor-made for robots. Professor Gary Becker's "high sanctions infrequently applied" approach seems unfair in many human contexts because it can have widely varied interpersonal effects: even if we get equal deterrence from a 100 percent chance of a year in prison or a 10 percent chance of ten years in prison, the lottery system that punishes a few very harshly seems intuitively unfair. We want our laws to protect both victims *and* wrongdoers against some forms of moral bad luck (whereas Becker's approach exacerbates it). But robots will internalize the probability of punishment as well as its magnitude, so we may be able to encourage efficient behavior without worrying about treating all robots equitably.

Even if we decide to heed Becker's advice, getting the numbers right presumes that we have a good estimate of the proportion of torts committed by robots that go undetected. That's tough to do, especially for newly introduced technologies. Maybe society will instead be able to force corporations to internalize their costs through nonlegal mechanisms – for example, by voting with their wallets when a company's robots engage in misconduct. In the era of big data and even bigger trade secrets, structural asymmetries often prevent meaningful public engagement with the data and software critical to measuring and understanding the behavior of complex machines.

Current trends in AI appear likely to only exacerbate this problem. As Bryce Goodman and Seth Flaxman observe, systems capable of achieving the richest predictive results tend to do so through

the use of aggregation, averaging, or multilayered techniques which, in turn, make it difficult to determine the exact features that play the largest predictive role (GOODMAN; FLAXMAN, 2017). Thus, even more so than yesteryear's AIs, understanding how modern robots arrive at a given decision can be prohibitively difficult, if not technically impossible. As a result, potentially unlawful or defective decision-making within such systems can often only be demonstrated in hindsight, after measuring the unevenly distributed outcomes once they have already occurred. And as systems get more complex, maybe not even then.

The risk presented by this combination of factors is not so much that corporations will intentionally build bad robots in order to eke out extra profits, but that "[bad] effects [will] simply happen, without public understanding or deliberation, led by technology companies and governments that are yet to understand the broader implications of their technologies once they are released into complex social systems." (see CAMPOLO et al, 2017, *36). Indeed, much of the misconduct that tomorrow's designers, policymakers, and watchdogs must guard against might not be intentional at all. Self--learning machines may develop algorithms that take into account factors we may not want them to, like race or economic status. But on some occasions, taking precisely those factors into account will actually get us to the ultimate result of interest.

For this reason, we think AI transparency is no panacea. Transparency is a desirable goal in the abstract. But it may inherently be at odds with the benefits of certain robotics applications. We may be able to find out *what* an AI system did.

Are we right to be bothered by this? Should we have a right to understand the mens rea of robots? Or to impute explanations so we can appropriately channel opprobrium? Our punitive and deterrence remedies are based on identifying and weeding out bad behavior.

The search for that bad behavior is much of what drives the "intuitive appeal of explainable machines." (see SELBST; BAROCAS, 2018, p. 1.126-1.129). But our intuitions may not always serve us well. The question is whether the demand for an explanation is actually serving legitimate purposes (Preventing Skynet? Stopping discrimination?) or just making us feel that we're the ones in charge. The punitive and equitable monetary side of remedies law wants to understand the "why" question because we want to assign blame. But that might not be a meaningful question when applied to a robot.

C.3 INHUMAN, ALL TOO INHUMAN

a) Punishing robots for responding to punishment. Even economic forms of deterrence – both legal and extralegal – will look different than they currently do when people or corporations are being deterred. Deterrence of people often takes advantage of cognitive biases and risk aversion. People don't want to go to jail, for instance, so they will avoid conduct that might lead to that result. But robots can be deterred only to the extent that their algorithms are modified to include external sanctions as part of the risk-reward calculus. Once more, we might view this as a good thing – the ultimate triumph of a rational law and economics calculus of decision-making. But humans who interact with robots may demand a noneconomic form of moral justice even from entities that lack the human capacity to understand the wrongfulness of their actions (a fact that anyone who has ever hit a malfunctioning device in frustration can understand).

Indeed, the sheer rationality of robot decision-making may itself provoke the ire of humans. Any economist will tell you that the optimal number of deaths from many socially beneficial activities is more than zero. Were it otherwise, our cars would never go more than five miles per hour. Indeed, we would rarely leave our homes at all.

Effective deterrence of robots requires that we calculate the costs of harm caused by the robots interacting with the world. If we want a robot to take optimal care, we need it to figure out not just how likely a particular harm is but how it should weight the occurrence of that harm. The social cost of running over a child in a crosswalk is high. But it isn't infinite (GEISTFELD, 2001, p. 125-126).

Even today, we deal with those costs in remedies law unevenly. The effective statistical price of a human life in court decisions is all over the map. The calculation is generally done ad hoc and after the fact. That allows us to avoid explicitly discussing politically fraught concepts that can lead to accusations of "trading lives for cash." (see generally, SUNSTEIN, 2004, p. 205) And it may work acceptably for humans because we have instinctive reactions against injuring others that make deterrence less important. But in many instances, robots will need to quantify the value we put on a life if they are to modify their behavior at all. Accordingly, the companies that make robots will have to figure out how much they value human life, *and they will have to write it down in the algorithm for all to see* (at least after extensive discovery).

The problem is that people strongly resist the idea of actually making this calculus explicit. They oppose the seemingly callous idea of putting a monetary value on a human life, and juries punish companies that make explicit the very cost-benefit calculations that economists want them to make (see, for example, SUNSTEIN, 2004, p. 205). Human instincts in this direction help explain why we punish intentional conduct more harshly than negligent conduct. A deliberate decision to run over a pedestrian strikes us as worse than hitting one by accident because you weren't paying attention. Our assumption is that if you acted deliberately, you could have chosen not to cause the harm, thereby making you a bad actor who needs to modify your behavior. But that assumption often operates even when causing that harm was the socially responsible thing to do, or at least was justified from a cost-benefit perspective.

b) Punishment as catharsis: punching robots. Punishment may serve other, nonmonetary purposes as well. We punish, for instance, to channel social opprobrium. That can set norms by sending a message about the sorts of things we won't tolerate as a society. And it may also make us feel better. We have victim allocution in court for good reason, after all. It may provide useful information to courts. But it also helps people to grieve and to feel their story has been heard.

Our instinct to punish is likely to extend to robots. We may want, as Professor Mulligan puts it, to punch a robot that has done us wrong. Certainly people punch or smash inanimate objects all the time. Juries might similarly want to punish a robot, not to create optimal cost internalization but because it makes the jury and the victim feel better (see ABBOTT; SARCH, 2019, *17-19).

That kind of expressive punishment may also stem from the fact that much human behavior is regulated by social sanction, not just law. Aggressively signaling social displeasure doesn't just make us feel better; it sends an object lesson to others about what is not acceptable behavior. Our instinct makes us want to send that lesson to robots too.

It's already quite easy to think of robots as humans. We naturally anthropomorphize (see CALO, 2015, p. 513, 545-549). That instinct is likely to get stronger over time, as companies increasingly deploy "social robots" that intentionally pull on these strings. Humans will expect humanlike robots to act, well, human. And we may be surprised, even angry, when they don't. Our instinct may increasingly be to punish humanoid robots as we would a person – even if, from an economic perspective, it's silly. Making us feel better may be an end unto itself. But hopefully there is a way to do it that doesn't involve wanton destruction of or damage to robots.

D ORDERING ROBOTS TO BEHAVE

All these problems with monetary remedies as deterrents seem to point in the direction of using injunctive relief more with robots than we currently do with people. Rather than trying to encourage robot designers to build in correctly priced algorithms to induce efficient care, wouldn't it be easier just to tell the robot what to do – and what not to do?

D.1 BE CAREFUL WHAT YOU WISH FOR

First, the good news: injunctions against robots might be simpler than against people or corporations because they can be enforced with code. A court can order a robot, say, not to take race into account in processing an algorithm. Likewise, it can order a self-driving car not to exceed the speed limit. Someone will have to translate that injunction, written in legalese, into code the robot can understand. But once they do, the robot will obey the injunction. This virtual guarantee of compliance seems like a significant advantage over existing injunctions. It is often much harder to coerce people (and especially groups of people in corporations) to comply with similar court orders – even when the consequences are dire.

But once again, not so fast. As the adage goes (and as legions of genies in bottles have taught us): be careful what you wish for. Automatic, unthinking compliance with an injunction is a good idea only if we're quite confident that the injunction itself is a good idea. Now, obviously the court thinks the injunction improves the world. Otherwise, it wouldn't issue it. But the fact that injunctions against people aren't self-enforcing offers some potential breathing room for parties and courts to add a dose of common sense when circumstances change. This is a common problem in law. It's a major reason we have standards rather than rules in many cases. And it's

the reason that even when we do have rules, we don't enforce them perfectly. To a person (and even to a police officer), "don't exceed the speed limit" implicitly means "don't exceed the speed limit unless you're rushing someone to the emergency room or it would be unsafe not to speed."

Try telling that to a robot, though. Machines, unlike at least some humans, lack common sense. They operate according to their instructions – no more, no less. If you mean "don't cross the double yellow line unless you need to swerve out of the lane to avoid running over a kid" you need to say that. Meanwhile, autonomous vehicles should probably avoid adults too, so better put that in the algorithm. … And maybe dogs. ... And deer and squirrels, too. Or maybe not – crossing into oncoming traffic is dangerous, so while we might do it to avoid hitting a kid even if it raises the risk of a head-on collision, we shouldn't do it to avoid a squirrel unless the risk of a head-on collision seems low. If you want the self-driving car to do all that, you need to tell it so. That's hard. It's more plausible to give each outcome weights – killing squirrels is bad, but head-on collisions are much worse, and killing a kid is (Probably? Maybe?) worse still. But then we're back to deterrence and cost internalization, not injunctions.

Further, even if we can specify the outcome we want with sufficient precision in an injunction, we need to be extremely careful about the permissible means a robot can use to achieve that result. Think back to our example from the Introduction. The drone did exactly what we told it to. The problem is that we weren't sufficiently clear in communicating what we wanted it to do.

To issue an effective injunction that causes a robot to do what we want it to do (and nothing else) requires both extreme foresight and extreme precision in drafting it. If injunctions are to work at all, courts will have to spend a lot more time thinking about exactly

what they want to happen and all the possible circumstances that could arise. If past experience is any indication, courts are unlikely to do it very well. That's not a knock on courts. Rather, the problem is twofold: words are notoriously bad at conveying our intended meaning, and people are notoriously bad at predicting the future. Coders, for their part, aren't known for their deep understanding of the law, and so we should expect errors in translation even if the injunction is flawlessly written. And if we fall into any of these traps, the consequences of drafting the injunction incompletely may be quite severe.

D.2 "WHAT DO YOU MEAN YOU CAN'T?!"

Courts that nonetheless persist in ordering robots not to do something may run into a second, more surprising problem: it may not be simple or even possible to comply with the injunction. Just as robots don't have money, they also don't read and implement court opinions.[10] And they aren't likely to be a party to the case in any event. Enjoining a robot, in other words, really means ordering someone else to implement code that changes the behavior of the robot.

The most likely party to face such an injunction is the owner of the robot. The owner is the one who will likely have been determined to have violated the law, say by using a discriminatory algorithm in a police-profiling decision or operating a self-driving car that has behaved unsafely. But most owners won't have the technical ability, and perhaps not even the right, to modify the algorithm their robot runs. The most a court could order may be that they ask the vendor who supplied the robot to make the change, or perhaps to take the robot off the market as long as it doesn't comply with the injunction.

[10] Well, some do.

Even if the developer is a party to the case, perhaps on a design defect theory, the self-learning nature of many modern robots makes simply changing the algorithm more complicated still. A court may, for instance, order the designer of a robot that makes predictions about recidivism for parole boards not to take race into account. But that assumes that the robot is simply doing what it was originally programmed to do. That may be less and less common as machine learning proliferates. Ordering a robot to "unlearn" something it has learned through a learning algorithm is much less straightforward than ordering it to include or not include a particular function in its algorithm. Depending on how the robot learns, it might not even be possible.

Life gets easier if courts can control what training information is fed to robots in the first place. At the extremes, a court might order a company to take badly trained robots out of service and to train new ones from scratch. But as the example in the Introduction indicates, the effects of training material on robots are not always predictable. And the results of training are themselves unpredictable, so even controlling the training dataset is no guarantee that a robot, once trained, will behave as the court wants it to.

Further, the future may bring robots that are not only trained in complicated ways but that train themselves in ways we do not understand and cannot replicate. Ordering such a robot to produce or not produce a particular result, or even to consider or not consider a particular factor, may be futile. Courts are used to telling people to do something and having them do it. They may have little patience for the uncertainties of machine learning systems. And they are quite likely to have even less patience with lawyers who tell them their "client" can't comply with the court's order.

D.3 UNINTENDED CONSEQUENCES

Even when the injunction is clear and codifiable, ordering a robot to change how it "thinks" is likely to have unintended consequences. Consider two examples.

(1) We don't want self-driving cars to hit pedestrians. But just brute-forcing that result might lead to other problems, from taking crowded freeways instead of less-crowded surface streets to running into other cars. Some of those consequences could be worse, either because a head-on collision kills more people than running over the pedestrian would or, more likely, because instructing the car to act in a certain way may cause it to avoid a very small chance of killing a pedestrian by avoiding surface streets altogether (even though the collective cost of traffic jams might be quite great). This is a version of the same problem we saw in damages: we need to assign a cost to various outcomes if we want an algorithm to weigh the alternatives. But here the injunction effectively sets the cost as infinite. That's fine if there really is nothing to balance on the other side. But that will rarely be true.

(2) The case against algorithmic bias seems one of the strongest, and easiest to enjoin, cases. And if that bias results simply from a bad training set, it may be straightforward to fix. But if an algorithm takes account of a prohibited variable like race, gender, or religion *because that variable matters in the data*, simply prohibiting consideration of that relevant information can have unanticipated consequences. One possible consequence is that we make the algorithm worse at its job. We might be fine as a society with a certain amount of that in exchange for the moral clarity that comes with not risking discriminating against minorities. But where people are in fact different, insisting on treating them alike can itself be a form of discrimination. Being male, for example, is an extremely strong predictor of criminality. Men commit many more

crimes than women (see LOESCHE, 2017), and male offenders are much more likely to reoffend (ALPER; DUROSE; MARKMAN, 2018, *6). We suspect police and judges know this and take it into account, consciously or unconsciously, in their arrest, charging, and sentencing decisions, though they would never say so out loud. But a robot won't conceal what it's doing. A court that confronts such a robot is likely to order it not to take gender into account, since doing so seems a rather obvious constitutional violation. But it turns out that if you order pretrial sentencing algorithms to ignore gender entirely, you end up discriminating against women, since they get lumped in with the heightened risks of recidivism that men pose.

Ordering a robot not to violate the law can lead to additional legal difficulties when injunctions are directed against discrete subsystems within larger robotics systems. These types of injunctions seem likeliest to be granted against newly introduced subsystems within a tried and true application – given that older systems will, by definition, have a longer track record of success. Not only could targeting one component of a larger system change it in unpredictable and often undesirable ways, doing so could also discourage innovation. With the field of AI improving by leaps and bounds, maybe we should be less protective of tried-and-true approaches and more willing to experiment. Even though some of those experiments will fail, the overall arc is likely to bend toward better systems than we have now. But we won't get there if courts are too quick to shut down new systems while leaving established but imperfect procedures in place. If the alternative to a flawed predictive policing algorithm is the gut instincts of a large number of cops, some of whom are overtly racist and others of whom are subconsciously biased, we might be better off with the robots after all.

II LESSONS FOR PUNISHING HUMANS

Robots present a number of challenges to courts imposing remedies on robotic and AI defendants. And we've only scratched the surface in this Essay. Working through these challenges is valuable and important in its own right. But doing so also teaches us some things about the law of remedies as it currently applies to people and corporations.

First, much of remedies, like much of law, is preoccupied with fault – identifying wrongdoers and treating them differently. There may be good reasons for that, both within the legal system and in society as a whole. But it works better in some types of cases than in others. Our preoccupation with blame motivates many remedies, particularly monetary equitable relief. This preoccupation distorts damage awards, particularly when something really bad happens and there is not an obvious culprit. It also applies poorly to corporations, which don't really have a unitary purpose in the way a person might. It's also costly, requiring us to assess blame in traffic cases that could otherwise be resolved more easily if we didn't have to evaluate witness credibility. A fault-based legal system doesn't work particularly well in a world of robots. But perhaps the problem is bigger than that: it might not work well in a world of multinational corporations either. We should look for opportunities to avoid deciding fault, particularly when human behavior is not the primary issue in a legal case.

A second lesson is the extent to which our legal remedies, while nominally about compensation, actually serve other purposes, particularly retribution. Remedies law can be described as being about "what you get when you win." But decades of personal experience litigating cases have reinforced the important lesson that what plaintiffs want is quite often something the legal system isn't prepared to give. They may want to be heard, they may want justice

to be done, or they may want to send a message to the defendant or to others. Often what they want – closure, or for the wrong to be undone – is something the system not only can't give them, but that the process of a lawsuit actually makes worse. The disconnect between what plaintiffs want and what the law can give them skews remedies law in various ways. Some do no harm: awards of nominal damages or injunctions that vindicate a position while not really changing the status quo. But we often do the legal equivalent of punching robots – punishing people to make ourselves feel better, even as we frequently deny compensation for real injuries. It's just that it's easier to see when it's a robot you're punching.

A final lesson is that our legal system sweeps some hard problems under the rug. We don't tell the world how much a human life is worth. We make judgments on that issue every day, but we do them haphazardly and indirectly, often while denying we are doing any such thing. We make compromises and bargains in the jury room, awarding damages that don't reflect the actual injury the law is intended to redress but some other, perhaps impermissible consideration. And we make judgments about people and situations in- and outside of court without articulating a reason for it, and often in circumstances in which we either couldn't articulate that decision--making process or in which doing so would make it clear we were violating the law. We swerve our car on reflex or instinct, sometimes avoiding danger but sometimes making things worse. We don't do that because of a rational cost-benefit calculus, but in a split-second judgment based on imperfect information. Police decide whether to stop a car, and judges whether to grant bail, based on experience, instinct, and bias as much as on cold, hard data.

Robots expose those hidden aspects of our legal system and our society. A robot can't make an instinctive judgment about the value of a human life, or about the safety of swerving to avoid

a squirrel, or about the likelihood of female convicts reoffending compared to their male counterparts. If robots have to make those decisions – and they will, just as people do – *they will have to show their work*. And showing that work will, at times, expose the tolerances and affordances our legal system currently ignores. That might be a good thing, ferreting out our racism, unequal treatment, and sloppy economic thinking in the valuation of life and property. Or it might be a bad thing, particularly if we have to confront our failings but can't actually do away with them. It's probably both. But whatever one thinks about it, robots make explicit many decisions our legal system and our society have long decided not to think or talk about. For that, if nothing else, remedies for robots deserve serious attention.

Failing to recognize this fact could result in significant unintended consequences – inadvertently encouraging the wrong behaviors, or even rendering our most important remedial mechanisms functionally irrelevant. Robotics will require some fundamental rethinking of what remedies we award and why. That rethinking, in turn, will expose a host of fraught legal and ethical issues that affect not just robots but people, too. Indeed, one of the most pressing challenges raised by the technology is its tendency to reveal the tradeoffs between societal, economic, and legal values that many of us, today, make without deeply appreciating the downstream consequences.

In a coming age where robots play an increasing role in human lives, ensuring that our remedies rules both account for these consequences *and* incentivize the right ones will require care and imagination. We need a law of remedies for robots. But in the final analysis, remedies for robots may also end up being remedies for all of us.

REFERENCES

ABBOT, Ryan; SARCH, Alex. Punish artificial inteligence: legal fiction or science fiction. In: **University of California Davis Law Review**, v. 53, page 323-384, November 2019.

ALPER, Mariel; DUROSE, Matthew R.; MARKMAN, Joshua. **2018 update on prisoner recidivism:** a 9-year follow-up period (2005–2014). Washington: U.S. Department of Justice Office of Justice, Programs, Bureau of Justice Statistics, May 2018.

BEZEMEK, Christoph. Bad for good: perspectives on law and force. In: BEZEMEK, Christoph; LADAVAC, Nicoletta (Ed.). **The force of law reaffirmed**: Frederick Schauer meets the critics. Switzerland: Springer, 2016.

CALO, Ryan. Robotics and the lessons of cyberlaw. In: **California Law Review**, Berkeley, Ed. University of California, Berkeley, v. 103, Issue 3, p. 513-564, 2015.

CAMPOLO, Alex et al. **AI Now 2017 Report**. New York: AI Now Institute at New York University, 2017.

GEISTFELD, Mark. Reconciling cost-benefit analysis with the principle that safety matters more than money. In: **New York University Law Review**, New York, Ed. New York University, v. 76, n. 1, p. 114-189, 2001.

GOODMAN, Bryce; FLAXMAN, Seth. European Union Regulations on algorithmic decision-making and a "right to explanation". In: ICML Workshop on Human Interpretability in Machine Learning, 6, New York, NY, June 23, 2016. **Proceedings…** p. 26-30, 2017.

HOLMES JR, Oliver Wendell. The path of the Law. In: **Harvard Law Review**, Cambridge (Massachusetts), Ed. University Harvard, v. 10, n. 8, p. 457-478, 1897.

LAYCOCK, Douglas. **Modern American remedies**: cases and materials. 2020. (4th ed. Aspen: Wolters Kluwer, 2010).

LEMLEY, Mark. The fruit of the poisonous tree in IP law. In: **Iowa Law Review**, Iowa City, Ed. University of Iowa, v. 103, Issue 1, p. 245-269, 2017.

LOESCHE, Dyfed. The prison gender gap. In: **Statista**, New York, Oct 23, 2017.

VINCENT, James. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day In: **The Verge**, Mar 24, 2016.

ROSS, H. Laurence. **Settled out of Court**: the social process of insurance claims adjustment. 2nd ed. Chicago: Aldine, 1980.

SELBST, Andrew D.; BAROCAS, Solon. The intuitive appeal of explainable machines. In: **Fordham Law Review**, Ed. Fordham University School of Law. v. 87, Issue 3, 2018. p. 1.085-1.139.

SCIENCE diction: the origin of the word ''robot''. In: **NPR: 50 Hear Every Voice**, Apr. 22, 2011.

SUNSTEIN, Cass R. Lives, life-years, and willingness to pay. In: **Columbia Law Review**, Chicago, Ed. University of Chicago Law School, v. 104, Issue 2, p. 105-252, march 2004.

Recebido em: 25-6-2021
Aprovado em: 26-11-2021