Remedies for Robots[1]

Mark A. Lemley[2] & Bryan Casey[3]

Engineers training an artificially-intelligent self-flying drone were perplexed. They were trying to get the drone to stay within a circle and to head towards the center of that circle. Things were going well for a while. The drone received positive reinforcement for successful flights, and it was learning to fly towards the middle of the circle more quickly and accurately. Then, suddenly, things changed. When the drone was near the edge of the circle, it would turn and fly *away* from the center, leaving the circle.

What went wrong? After a long time puzzling over the problem, the designers realized that when the drone left the circle during the test, they shut it down, and someone picked it up and carried it back into the circle to start the test over again. The learning algorithm in the drone had figured out – correctly – that if it was sufficiently far from the center, the easiest way for it to get back to the middle was to leave the circle. From the drone's perspective, when it left the circle altogether, it was magically teleported back to the middle of the circle. The drone had found a short cut.[4] It had complied with the rules it was given, but it had done so in a way that subverted the trainer's intent.

What happens when artificially intelligent robots misbehave, as the drone did here? The question is not just hypothetical. As robots and artificial intelligence (AI) increasingly integrate into our society, they will do bad things. Sometimes they will cause harm because of a design or

[2] William H. Neukom Professor, Stanford Law School; partner, Durie Tangri LLP.

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

[4] This example comes from a presentation at the June 2014 Stanford Ecommerce Best Practices Conference. As far as I know it has not been previously described in print.

implementation defect: we should have programmed the self-driving car to recognize a graffiti-covered stop sign but we failed to do so. Sometimes they will cause harm because doing so is a necessary byproduct of the intended operation of the machine. Cars kill lots 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—but one that, nonetheless, does not sit well with us. And sometimes, as with our drone, they will do unexpected things for reasons that doubtless have their own logic but which we can't understand or predict.

These new technologies present a number of interesting substantive law questions, from predictability to transparency to how we should assess liability and judgment in robots and AI systems. A growing body of scholarship is beginning to address these questions.[5] Together, those scholars are beginning to think about the substantive law that should govern robots and AI.

Our focus here is different: what remedies can and should the law provide once a robot has caused harm. The law of remedies is trans-substantive. Where substantive law defines who wins legal disputes, remedies law asks, "What do I get when I win?" Remedies law is sometimes designed to make plaintiffs whole by restoring them to the condition they would have been in but for the wrong. But remedies law also contains many elements of moral judgment, punishment, and deterrence. For instance, the law will often act to deprive the defendant of its gain even if the result is a windfall to the plaintiff, because we think it is unfair to let the defendant keep that gain. And often we want to order the defendant to do (or stop doing) something unlawful or harmful.

Each of those goals runs into difficulties when the bad actor is neither a person nor a corporation but a robot. We might order a robot (or, more realistically, the designer or owner of

[5] Lots of cites here. Calo, Surden, Darling, algorithm discrimination papers, Abbott on ownership and obviousness, Casey

the robot) to pay for the damage it has caused, 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 to engage in certain conduct or refrain from other conduct. That is true particularly when the AI learns and modifies its decision-making over time, as the drone in the opening example did. And it is harder still to think sensibly about punishing a robot or depriving it of benefits it has gained unfairly.

One way to avoid this problem is to try to move responsibility up the chain of command from a robot to its human or corporate masters – either the designers or the owners who employ it. But that too comes with its problems. Robot decision-making is increasingly likely to be based on learning algorithms. The developers – and certainly the users – of those algorithms won't necessarily know or be able to deterministically control the inputs into the robot's decisionmaking process. And it won't be easy to figure out how to allocate responsibility between them. If the goal is to shape behavior or discourage bad behavior, punishing owners or designers for the behavior of robots may not make sense for the simple reason that their owners didn't act wrongfully in any meaningful way. 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 not to engage in certain behavior won't be trivial in many cases. Ordering it to take affirmative acts will be even more problematic.

In this paper, we begin to think about how we might design a system of remedies for robots. We might have to focus less attention on moral guilt and more on a no-fault liability system (or at least one that defines fault differently) to compensate plaintiffs. But paying for injury solves only part of the problem. Often we want to compel defendants to do (or not do) something in order to prevent injury. Injunctions, punitive damages, and even remedies like disgorgement are all aimed,

directly or indirectly, at modifying or deterring behavior. But ordering a robot to do something is different than ordering a person to do it – sometimes easier, sometimes harder. And deterring robot misbehavior is going to look very different than deterring people. Deterrence of people often takes advantage of cognitive biases and risk aversion. People don't want to go to jail, for instance, so they tend to 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. Perhaps we need a "robot death penalty" – shutting down dangerous robots as a sort of specific deterrence against bad behavior.

Finally, remedies law also has an expressive component. We sometimes grant punitive damages – or disgorge ill-gotten gains – to show our displeasure with you. If our goal is just to feel better, perhaps we could punish robots just for the sake of punishing them. Christina Mulligan half-jokingly suggests you should have the right to punch a robot.[6] But if our goal is to send a signal, robots will require us to rethink the nature of that signal.

In Part I, we discuss the development of robots and learning AIs and the sorts of robot wrongdoing that will increasingly draw the attention of the legal system. In Part II we outline the basic principles of remedies law and consider how those remedies will work – or not work – when applied to robots and AIs. In Part III we consider how we might remake remedies law with robots in mind.

[6] Christina Mulligan, *Revenge Against Robots*