**Liability Law for Present and Future Robotics Technology**

Trevor N. White and Seth D. Baum
Global Catastrophic Risk Institute
http://gcrinstitute.org * http://sethbaum.com

Published in *Robot Ethics 2.0*, edited by Patrick Lin, George Bekey, Keith Abney, and Ryan Jenkins, Oxford University Press, 2017, pages 66-79.
This version 10 October 2017

**Abstract**
Advances in robotics technology are causing major changes in manufacturing, transportation, medicine, and a number of other sectors. While many of these changes are beneficial, there will inevitably be some harms. Who or what is liable when a robot causes harm? This paper addresses how liability law can and should account for robots, including robots that exist today and robots that potentially could be built at some point in the near or distant future. Already, robots have been implicated in a variety of harms. However, current and near-future robots pose no significant challenge for liability law: they can be readily handled with existing liability law or minor variations thereof. We show this through examples from medical technology, drones, and consumer robotics. A greater challenge will arise if it becomes possible to build robots that merit legal personhood and thus can be held liable. Liability law for robot persons could draw on certain precedents, such as animal liability. However, legal innovations will be needed, in particular for determining which robots merit legal personhood. Finally, a major challenge comes from the possibility of future robots that could cause major global catastrophe. As with other global catastrophic risks, liability law could not apply, because there would be no post-catastrophe legal system to impose liability. Instead, law must be based on pre-catastrophe precautionary measures.

**Introduction**
In June 2005, a surgical robot at a hospital in Philadelphia malfunctioned during a prostate surgery, possibly injuring the patient.[1] In June 2015, a worker at a Volkswagen plant in Germany was crushed to death by a robot that was part of the assembly process.[2] In November 2015, a self-driving car in California made a complete stop at an intersection and then was hit by a car with a human driver, apparently because the self-driving car followed traffic law but not traffic norms.[3] These are just some of the ways that robots are already implicated in harms. As robots become more sophisticated and more widely adopted, the potential for harm will get even larger. Robots even show potential for causing harm at massive catastrophic scales.

How should robot harms be governed? In general, liability law governs harms in which someone or something else is responsible. Liability law is used to punish those who have caused harms, particularly those that could have and should have been avoided. The threat of punishment further serves to discourage those who could cause harm. Liability law is thus an important legal tool for serving justice and advancing the general welfare of society and its members. The value of liability law holds for robotics just as it does for any other harm-causing technology.

But robots are not just any other technology. Robots are (or at least can be) intelligent, autonomous actors moving about in the physical world. They can cause harms through actions that they choose to make, actions that no human told them to make and, indeed, that may surprise

their human creators. Perhaps robots should be liable for their harms. This is a historic moment: humans creating technology that could potentially be liable for its own actions. Furthermore, robots can have the strength of industrial machinery and the intelligence of advanced computer systems. Robots can also be mass produced and connected to each other and to other technological systems. This creates the potential for robots to cause unusually great harm.

This paper addresses how liability law can and should account for robots, including robots that exist today and robots that potentially could be built at some point in the near or distant future. Three types of cases are distinguished, each with very different implications. First are cases in which some human party is liable, such as the manufacturer or the human using the robot. These cases pose no novel challenges for liability law: they are handled the same way as with other technologies in comparable circumstances. Second are cases in which the robot itself is liable. These cases require dramatic revision to liability law, including standards to assess when robots can be held liable and principles for dividing liability between the robot and the humans who designed, built, and used it. Third are cases in which the robot poses a major catastrophic risk. These cases merit separate attention because a sufficiently large catastrophe would destroy the legal system and thus the potential to hold anyone or anything liable.

The three types of cases differ across two dimensions as shown in Figure 1. One dimension is the robot's degree of legal personhood, meaning the extent to which a robot shows attributes that qualify it for independent standing in a court of law. As we discuss, a robot can be held liable in the eyes of the law to the extent that it merits legal personhood. The other dimension shows the size of the harm the robot causes. Harms of extreme severity cannot be handled by liability law. However, there is no strict distinction between the three cases. Instead, there is a continuum, as shown by the regions in which a robot can have partial liability or more-than-human liability and in which liability law works to a limited extent.

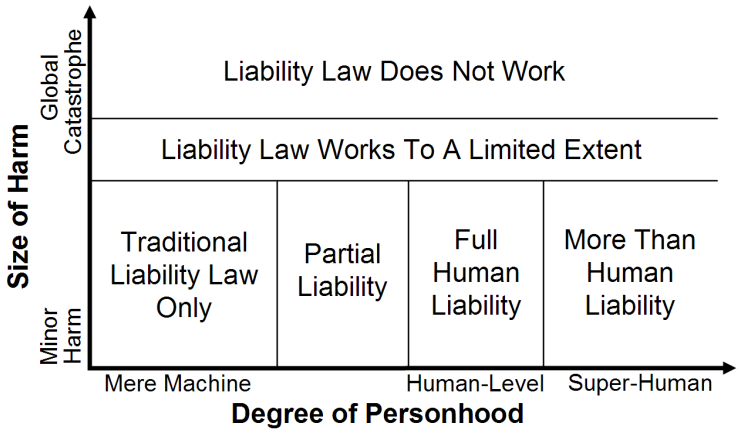


**Figure 1.** Classification scheme for the applicability of liability law to various sizes of harms caused by various types of robots.

## I - A Human Party Is Liable

In a detailed study of robot law, Weaver (2014, 21–27) identifies four types of parties that could be liable for harm caused by a robot: (1) people who were using the robot or overseeing its use; (2) other people who were not using the robot but otherwise came into contact with it, which can include people harmed by the robot; (3) some party involved in the robot's production and distribution, such as the company that manufactured the robot; or (4) the robot itself.

For the first three types of parties, liability applies the same as for other technologies. A surgical robot, for example, can be misused by the surgeon (type 1), bumped into by a hospital visitor who wandered into a restricted area (type 2), or poorly built by the manufacturer (type 3). The same situations can also arise for other, non-robotic medical technologies. In each case, the application of liability law is straightforward. Or rather, to the extent that the application of liability law is not straightforward, the challenges faced are familiar. The fourth type—when the robot is liable—is the only one that poses novel challenges for liability law.

To see this, consider one of the thornier cases of robot liability, that of lethal autonomous weapon systems (LAWSs). These are weapons that decide for themselves whom to kill. Sparrow (2007) argues that there could be no one liable for certain LAWS harms—for example, if a LAWS decides to kill civilians or soldiers who have surrendered. A sufficiently autonomous LAWS could make its own decisions, regardless of how humans designed and deployed it. In this case, Sparrow argues, it would be unfair to hold the designer or deployer liable (or the manufacturer or other human parties). It might further be inappropriate to hold the robot itself liable, if it is not sufficiently advanced in legally relevant ways (more on this in Section II). In this case, who or what to hold liable is ambiguous.

This ambiguous liability is indeed a challenge, but it is a familiar one. In the military context, precedents include child soldiers (Sparrow 2007, 73–74) and landmines (Hammond 2015, note 62). Child soldiers can make their own decisions, disobey orders, and cause harm in the process. Landmines can linger long after a conflict, making it difficult or impossible to identify who is responsible for their placement. In both cases, it can be difficult or perhaps impossible to determine who is liable. So too for LAWSs. This ambiguous liability can be a reason to avoid or even ban the use of child soldiers, landmines, and LAWSs in armed conflict.

Regardless, even for this relatively thorny case of robot liability, robotics technology raises no new challenges for liability law. In the United States, agencies such as the Department of Defense produce regulations on the use of LAWSs which are not dramatically different than for other weapons. Internationally, bodies like the UN's International Court of Justice could hold a state liable for authorizing drone strikes that caused excessive civilian casualties. Meanwhile, commercial drones can be regulated as other aircraft are now: by a combination of the FAA and corporate oversight by their creators (McFarland 2015). The handling of such relatively simple robots under liability law will thus be familiar if not straightforward.

The above LAWSs examples also resemble how liability law handles non-human animals, which has prompted proposals for robots to be given legal status similar to non-human animals (e.g., Kelley et al. 2010). Suppose someone gets a pet dog and then the dog bites someone, despite the owner trying to stop it. If this person had absolutely no idea the dog would bite someone, then she would not be liable for that bite. However, having seen the dog bite someone, she now knows the dog is a biter, and is now expected to exercise caution with it in the future. If the dog bites again, she can be liable. In legal terms, this is known as having *scienter*—knowledge of the potential harm.

Scienter could also apply to LAWSs or other robots that are not expected to cause certain harms. Once the robots are observed causing those harms, their owners or users could be liable for subsequent harms. For comparison, the Google Photos computer system raised controversy in 2015 when it mislabeled photographs of black people as "gorillas" (Hernandez 2015). No Google programmer instructed Photos to do this; it was a surprise, arising from the nature of Photos's algorithm. Google acted immediately to apologize and fix Photos. While it did not have scienter for the gorilla incident, it would for any subsequent offenses.[4] The same logic also

applies for LAWSs or other types of robots. Again, as long as a human party was responsible for it, a robot does not pose novel challenges to liability law.

Even if a human is ultimately liable, a robot could still be taken to court. This would occur, most likely, under *in rem* jurisdiction, in which the court treats an object of property as a party to a case when it cannot do so with a human owner. *In rem* cases include *United States v. Fifty-Three Electus Parrots* (1982), in which a human brought parrots from southeast Asia to the U.S. in violation of an animal import law, and *United States v. Forty Barrels & Twenty Kegs of Coca-Cola* (1916), in which the presence of caffeine in the beverage was at issue. In both cases, a human (or corporation) was ultimately considered liable, with the parrots and soda only serving as stand-ins. Robots could be taken to court in the same way, but they would not be considered liable except in a symbolic or proxy fashion. Again, since the robot is not ultimately liable, it poses no novel challenges to liability law.

This is not to say that such robots do not pose challenges to liability law—only that these are familiar challenges. Indeed, the nascent literature on robot liability identifies a range of challenges, including assigning liability when robots can be modified by users (Calo 2011), when they behave in surprising ways (Vladeck 2014), and when the complexity of robot systems makes it difficult to diagnose who is at fault (Funkhouser 2013; Gurney 2013). There are also concerns that liability laws could impede the adoption of socially beneficial robotics (e.g., Marchant and Lindor 2012; Wu 2016). However, these challenges all point to familiar solutions based in various ways of holding manufacturers, users, and other human parties liable. Fine tuning the details is an important and nontrivial task, but it is not a revolutionary one.

The familiar nature of typical robots to liability law is further seen in court cases in which robots have been implicated in harms (Calo 2016). An early case is *Brouse v. United States* (1949), in which two airplanes crashed, one of which was a US military plane that was using an early form of autopilot. The court rejected the US claim that it should not be liable because the plane was being controlled by the robotic autopilot; instead the court found that the human pilot in the plane is obligated to pay attention and avoid crashes. More recently, in *Ferguson v. Bombardier Services Corp.* (2007), another airplane crash may have been attributable to the autopilot system, in which case the court would have found the autopilot manufacturer liable, not the autopilot itself, but instead it found that the airline had improperly loaded the plane. (See Calo 2016 for further discussion of these and other cases.)

**II - Robot Liability**

If a robot can be held liable, then liability law faces some major challenges in terms of which robots to hold liable for which harms, and in terms of how to divide liability between the robot and its human designers, manufacturers, users, etc. In this section, we will argue that robots should be able to be held liable to the extent that they qualify for legal personhood. First, though, let us briefly consider some alternative perspectives.

One perspective is that, in an informal sense, any sort of object can be liable for a harm. The pollen in the air is liable for making you sneeze. The faulty gas pipe is liable for burning down your home. The earthquake is liable for destroying the bridge. This is not the sense of liability we address in this paper. Our focus is on legal liability, in which a party can be tried in court.

Another perspective comes from the notion that the law ultimately derives from what members of a society want it be. This is why laws are different in different jurisdictions and at different times. From this perspective, robots will be held liable whenever societies decide to hold them liable. There are difficult issues here, such as whether to give robots a say in if they

should be held liable.[5] Regardless, the fact that laws are products of societies need not end debate on what laws societies can and should have. To the contrary, it is incumbent upon members of society to have such debates.

Within human society, in the United States and many other countries, parties can be held liable for harms to the extent that they qualify as legal persons. Legal personhood is the ability to have legal rights and obligations, such as the ability to enter contracts, sue or be sued, and be held liable for one's actions. Legal liability thus follows directly from legal personhood. Normal adult humans are full legal persons and can be held liable for their actions across a wide range of circumstances. Children, the mentally disabled, and corporations have partial legal personhood, and in turn can be held liable across a narrower range of circumstances. Non-human animals generally do not have personhood, though this status has been contested, especially for non-human primates.[6]

The denial of legal personhood to non-human animals can be justified on grounds that they lack humans' cognitive sophistication and corresponding ability to participate in society. Such justification avoids charges of speciesism (a pro-human bias for no other reason than just happening to be human). However, the same justification implies that robots should merit legal personhood if they have similar capabilities as humans. As Hubbard (2011, 417) puts it, "Absent some strong justification, a denial of personhood to an entity with at least an equal capacity for personhood would be inconsistent and contrary to the egalitarian aspect of liberalism."[7]

The question of when robots can be liable thus becomes the question of when robots merit personhood. If robots merit personhood, then they can be held liable for harms they cause. Otherwise, they cannot be held liable, and instead liability must go to some human party, as is the case with non-human animals and other technologies or entities that can cause harm.

Hubbard proposes three criteria that a robot or other artificial intelligence should meet to merit personhood: (1) complex intellectual interaction skills, including the ability to communicate and learn from experience; (2) self-consciousness, including the ability to make one's own goals or life plan, and (3) community, meaning the ability to pursue mutual benefits within a group of persons. These three criteria, central to human concepts of personhood, may offer a reasonable standard for robot personhood. We will use these criteria for this paper while emphasizing that their definitude should be a matter of ongoing debate.

Do Hubbard's criteria also apply for liability? Perhaps not for the criterion of self-consciousness. The criterion makes sense for harms caused *to* a robot: only a conscious robot can experience harms as humans do.[8] This follows from, for example, classic utilitarianism, as in Bentham's line "The question is not, Can they reason? nor, Can they talk? but, Can they suffer?" However, the same logic does not apply to harms caused *by* a robot. Consider an advanced robot that meets all of Hubbard's criteria except that it lacks consciousness. Suppose the robot causes some harm—and, to be clear, the harm causes suffering to a human or to some other conscious person. Should the robot be held liable?

The answer to this may depend on society's foundational reasoning for liability. If liability exists mainly to discourage or deter the commission of harms, then consciousness is unnecessary. The robot should be punished so long as doing so discourages the commission of future harms. The entities that get discouraged here could include the robot, other similar robots, conscious robots, and even humans. It is quite conceivable that non-conscious robots could be punished with some sort of reduced reward or utility as per whatever reward/utility function they might have (Majot and Yampolskiy 2014). Specifically, they could be reprogrammed, deactivated or destroyed, or put in what is known as a "Box": digital solitary confinement

restricting an AI's ability to communicate or function (Corwin 2002; Yudkowsky 2002). To make this possible, however, such robots ought to be based (at least in part) on reinforcement learning or similar computing paradigms (except ones based on neural network algorithms, for reasons we explain later).

Alternatively, if liability exists mainly for retribution, to bring justice to whomever committed the harm, then consciousness could be necessary. Whether it is necessary depends on the purpose of the punishment. If the punishment aims to worsen the life of the liable party, so as to "balance things out," then consciousness seems necessary. It makes little sense to "worsen" the life of something that cannot experience the worsening. However, if the punishment aims to satisfy society's sense of justice, then consciousness may be unnecessary. Instead, it could be sufficient that members of society observe the punishment and see justice being served.[9] Whether the robot's consciousness would be necessary in this case would simply depend on whether society's sense of justice requires it to be conscious.

This potential exception regarding consciousness is a good example of partial liability as shown in Figure 1. The advanced, non-conscious robot can be held liable, but not in every case in which normal adult humans could. Specifically, the robot would not be held liable in certain cases where punishment is for retribution. Other limitations to a robot's capabilities could also reduce the extent of its liability. Such robots would be analogous to children and mentally disabled adult humans, who are similarly not held liable for as many cases as normal adult humans are. Robots of less sophistication along any of Hubbard's three criteria (or whatever other criteria are ultimately established) should be liable to a lesser extent than robots that meet the criteria in full.

What about robots of greater-than-human sophistication in Hubbard's three criteria? These would be robots with more advanced intellectual interaction skills, self-consciousness, or communal living ability. It is conceivable that such robots could exist—indeed, the idea dates back many decades (Good 1965). If they do come into existence, then by the above logic, they should be held to a *higher* liability standard than normal adult humans. Indeed, concepts such as negligence recognize human fallibility in many respects that a robot could surpass humans in, including reaction time, eyesight, and mental recall. The potential for holding robots to a higher standard of liability could offer one means of governing robots with greater-than-human capacities; more on this in Section III in the discussion of catastrophic risk.

Before turning to catastrophic risk, there is one additional aspect of robot liability to consider: the division of liability among the robot itself and other parties that influence the robot's actions. These other parties can include the robot's designer, its manufacturer, and any users or operators it may have. These parties are comparable to a human's parents and employers, though the comparison is imperfect due to basic differences between humans and robots.

One key difference is that robots are to a very large extent *designed*. Humans can be designed as well via genetic screening and related techniques, hence the term "designer baby." But designers have much more control over the eventual character of robots than they do for humans. This suggests that robot designers should hold more liability for robots' actions than human parents should for their children's actions. If robot designers know that certain designs tend to yield harmful robots, then a case can be made for holding the designers at least partially liable for harms caused by those robots, even if the robots merit legal personhood. Designers could be similarly liable for building robots using opaque algorithms, such as neural networks and related deep learning methods, in which it is difficult to predict in advance whether the robot

will cause harm. Those parties that commission the robot's design could be similarly liable. In court, the testimony of relevant industry experts would be valuable for proving whether any available, feasible safeguards to minimize such risks existed.

Another difference is that, at least for now, the production of robots is elective, whereas the birthing of humans is required for the continuity of society. Society cannot currently function without humans, but it can function without robots. This fact suggests some lenience for parents in order to encourage procreation, and to be stricter with robot designers in order to safely ease society's transition into an era in which humans and their robot creations coexist. Such a gradual transition seems especially warranted in light of potential robot catastrophe scenarios.

## III - Catastrophic Robot/AI Liability

"Catastrophe" has many meanings, many of which require no special legal attention. For example, a person's death is catastrophic for the deceased and her or his loved ones, yet the law is perfectly capable of addressing individual deaths caused by robots or AIs. However, a certain class of extreme catastrophe does merit special legal attention, due to its outsized severity and significance for human civilization. These are catastrophes that cause major, permanent harm to the entirety of global human civilization. Such catastrophes are commonly known as *global catastrophes* (Baum and Barrett 2016) or *existential catastrophes* (Bostrom 2013). Following Posner (2004), we will simply call them catastrophes.

A range of catastrophic risks exist, including global warming, nuclear war, a pandemic, and collision between Earth and a large asteroid or comet. Recently, a body of scholarship has built up analyzing the possibility of catastrophe from certain types of future AI. Much of the attention has gone to "superintelligent" AI that outsmart humanity and "achieve complete world domination" (Bostrom 2014, 78; see also Müller 2015). Such AI could harm humans through the use of robotics. Additionally, some experts believe that robotics could play an important role in the development of such AI (Baum et al. 2011).

Other catastrophe scenarios could also involve robotics. Robots could be used in the systems for launching nuclear weapons or for detecting incoming attacks, potentially resulting in unwanted nuclear wars.[10] They could be used in critical civil, transportation, or manufacturing infrastructure, contributing to a global systemic failure.[11] They could be used for geoengineering —the intentional manipulation of the global environment, such as to counteract global warming —and this could backfire, causing environmental catastrophe.[12] Robots could be used in establishing or maintaining an oppressive totalitarian world government.[13] Still further robot catastrophe scenarios may also be possible.

The enormous scale of the catastrophes in question creates profound moral and legal dilemmas. If the harm is permanent, it impacts members of all future generations, which could be immensely many people. Earth will remain habitable for at least a billion more years, and the galaxy and the universe for much longer (Baum 2016); the present generation thus contains just a tiny fraction of all people who could exist. The legal standing and representation of members of future generations is a difficult question (Tonn 1996; Wolfe 2008). If members of future generations are to be counted, then they can overwhelm the calculus. Despite this, present generations unilaterally make the decisions. There is thus a tension in how to balance the interests of present and future generations (Page 2003). A sufficiently large catastrophe raises similar issues even just within the context of the present generation. About seven billion humans live today; a catastrophe that risks killing all of them could be seven billion times larger than a

catastrophe that risks killing just one. One could justify enormous effort to reduce that risk regardless of future generations (Posner 2004).

Further complications come from the irreversible nature of these catastrophes. In a sense, every event is irreversible: if someone wears a blue shirt today, no one can ever change the fact that they wore a blue shirt today. Such events are irreversible only in a trivial sense: you can change what shirt you wear on subsequent days. Nontrivially irreversible events are more or less permanent: if that person should die today, then nothing[14] can bring that person back to life. At a larger scale, nontrivially irreversible effects exist for many ecological shifts and may also exist for the collapse of human civilization (Baum and Handoh 2014). The possibility of large and nontrivially irreversible harm creates a major reason to avoid taking certain risks. The precautionary principle is commonly invoked in this context, raising questions of just how cautious to be (Posner 2004; Sunstein 2006).

An irreversible AI catastrophe could be too large for liability law to handle. In the simplest case, if the catastrophe results in human extinction, then there would be no one remaining to hold liable. A catastrophe that leaves some survivors but sees the collapse of human civilization would lack the legal system needed for holding people liable. Alternatively, AI could cause a catastrophe in which everyone is still alive but they have become enslaved or otherwise harmed by the AI; in this case the pre-catastrophe human authorities would lack the power needed to hold those at fault liable. For smaller catastrophes, the legal system may exist to a limited extent (Figure 1). In this case, it may be possible to bring the liable parties to trial and/or punish them, but not as reliably or completely as is possible under normal circumstances. The closest possible example would be creating special international proceedings, like the Nuremberg Trials, to deal with the aftermath. Much like such war tribunals, though, these may do little to address the chaos' original cause. This would leave victims or society at large wasting time and resources on reliving a tragedy (McMorran 2013).

Hence, instead of liability, a precautionary approach could be used. This would set a default policy of disallowing any activity with any remote chance of causing catastrophe. It could further place the burden of proof on those who wish to conduct such activity, requiring them to demonstrate in advance that it could not cause catastrophe.[15] Trial-and-error would not be permitted, because a single error could cause major irreversible harm. This would likely be a significant impediment for AI research and development (at least for the subset of AI that poses catastrophic risk), which, like other fields of technology, is likely to make extensive use of trial and error. Indeed, some AI researchers recommend a trial-and-error approach, in which AIs are gradually trained to learn human values so that they will not cause catastrophe (Goertzel 2016). However, given the high stakes of AI catastrophe, perhaps these sorts of trial-and-error approaches should still be avoided.

It may be possible to use a novel liability scheme to assist with a catastrophe-avoiding precautionary approach. In a wide-ranging discussion of legal measures to avoid catastrophe from emerging technologies, Wilson (2013, 356) proposes "liability mechanisms to punish violators whether or not their activities cause any harm". In effect, people would be held liable not for causing catastrophe, but for taking actions that could cause catastrophe. This proposal could be a successful component of a precautionary approach to catastrophic risk and is worth ongoing consideration.

Taking the precautionary principle to the extreme can have undesirable consequences. All actions carry some risk. In some cases, it may be impossible to prove a robot does not have the potential to cause catastrophe. Therefore, requiring demonstrations of minimal risk prior to

performing actions would be paralyzing (Sunstein 2006). Furthermore, many actions can reduce some risks even while increasing others; requiring precaution due to concern about one risk can cause net harm to society by denying opportunities to decrease other risks (Wiener 2002). AI research and development can pose significant risks, but it can also help reduce other risks. For AI that poses catastrophic risk, net risk will be minimized when the AI research and development is expected to bring a net reduction in catastrophic risk (Baum 2014).

In summary, there are significant legal challenges raised by AI that poses catastrophic risk. Liability law, most critically, is of little help. Precautionary approaches can work instead, although care should be taken to avoid preventing AI from reducing different catastrophic risks. The legal challenges from AI that poses catastrophic risk is distinct from the challenges from other types of AI, but they are similar to the challenges from other catastrophic risks.

**Conclusion**

While robots benefit society in many ways, they also cause or are otherwise implicated in a variety of harms. The frequency and size of these harms is likely to increase as robots become more advanced and ubiquitous. Robots could even cause or contribute to a number of major global catastrophe scenarios. It is important for liability law to successfully govern these harms to the extent possible so that the harms are minimized and, when they do occur, that justice may be served.

For many robot harms, a human party is ultimately liable. For these harms, traditional liability law applies. A major challenge to liability law comes when robots could be liable. Such cases require legal personhood tests for robots to assess the extent to which they can be liable. One promising personhood test evaluates the robot's intellectual interaction skills, self-consciousness, and communal living ability. Depending on how a robot fares on a personhood test, it could have the same liability as, or less or more liability than, a normal adult human. A robot being liable does not preclude a human party also being liable. Indeed, robot designers should expect more liability for robot harms than would human parents, because robots are designed so much more extensively than human children are. Finally, for robots that pose catastrophic risk, liability law cannot be counted on and a precautionary approach is warranted.

People involved in the design, manufacture, and use of robots can limit their liability by choosing robots that reliably avoid harms. One potential way to improve reliability is to avoid computing paradigms such as neural nets that tend to result in surprising behaviors, or adapt these paradigms to make them less surprising (Huang and Xing 2002). Robot designs should be sufficiently transparent that the responsible human parties can, with reasonable confidence, determine in advance what harms could occur. They can then build safety restrictions into the robot or at least give warnings to robot users, as is common practice with other technologies. Robots should also go through rigorous safety testing before being placed into situations where they can cause harms. If robots cannot reliably avoid harms, then they probably should not be used in the first place.

These sorts of safety guidelines should be especially strict for robots that could contribute to major global catastrophe. A single catastrophe could permanently harm human civilization. It is thus crucial to avoid any catastrophe. Safety testing itself could be dangerous. This increases the value of transparent computing paradigms that let humans assess risks prior to building the robot. Legal measures must also take effect prior to the robot's build because there may be no legal system afterwards. Advanced robots may be less likely to cause catastrophe if they are designed

to be upstanding legal persons. But even then, some legal system would need to exist to hold them liable for what harms they cause.

As this paper illustrates, robot liability poses major new challenges to liability law. Meeting these challenges requires contributions from law, robotics, philosophy, risk analysis, and other fields. It is essential for humans with these various specialties to work together to build robot liability regimes that avoid harms while capturing the many benefits of robotics. The potential for harm is extremely large, making this an urgent task. We hope that humans and robots will coexist successfully and for mutual benefit in a community of responsible persons.

**Acknowledgments**

We thank Tony Barrett, Daniel Dewey, and Roman Yampolskiy for helpful comments on an earlier draft of this paper. All remaining errors or other shortcomings are ours alone. Work on this paper was funded in part by a grant from the Future of Life Institute. The views in this paper are those of the authors and do not necessarily reflect the views of the Global Catastrophic Risk Institute or the Future of Life Institute.

**References**

Barrett, Anthony M., Seth D. Baum, and Kelly R. Hostetler. 2013. "Analyzing and Reducing the Risks of Inadvertent Nuclear War Between the United States and Russia." *Science & Global Security* 21(2):106–133.

Barrett, Anthony M., and Seth D. Baum. 2016. "A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis." *Journal of Experimental & Theoretical Artificial Intelligence*. doi:10.1080/0952813X.2016.1186228.

Baum, Seth D. 2009. "Description, Prescription and the Choice of Discount Rates." *Ecological Economics* 69(1):197–205.

Baum, Seth D. 2010. "Universalist Ethics in Extraterrestrial Encounter." *Acta Astronautica* 66(3–4):617–23.

Baum, Seth D. 2014. "The Great Downside Dilemma for Risky Emerging Technologies." *Physica Scripta* 89(12). doi:10.1088/0031-8949/89/12/128004.

Baum, Seth D. 2016. "The Ethics of Outer Space: A Consequentialist Perspective." In *The Ethics of Space Exploration*, edited by James S.J. Schwartz and Tony Milligan. Berlin: Springer, forthcoming.

Baum, Seth D. and Anthony M. Barrett. 2016. "The Most Extreme Risks: Global Catastrophes." *The Gower Handbook of Extreme Risk*, edited by Vicki Bier. Farnham, United Kingdom: Gower, forthcoming.

Baum, Seth D., Timothy M. Maher, Jr., and Jacob Haqq-Misra. 2013. "Double Catastrophe: Intermittent Stratospheric Geoengineering Induced by Societal Collapse." *Environment, Systems and Decisions* 33(1):168–180.

Baum, Seth D. and Itsuki C. Handoh. 2014. "Integrating the Planetary Boundaries and Global Catastrophic Risk Paradigms." *Ecological Economics* 107:13–21.

Baum, Seth and Trevor White. 2015. "When Robots Kill." *The Guardian*, June 23. http://www.theguardian.com/science/political-science/2015/jul/23/when-robots-kill.

Bostrom, Nick 2013. "Existential Risk Prevention as Global Priority." *Global Policy* 4(1):15–31.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies* 78. Oxford: Oxford University Press.

Calo, Ryan. 2011. "Open Robotics." *Maryland Law Review* 70(3):571–613.

Calo, Ryan. 2016. “Robots in American Law.” *University of Washington School of Law Research Paper No. 2016-04*.

Caplan, Bryan. 2008. “The Totalitarian Threat.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan Ćirković, 504–19. Oxford: Oxford University Press.

Centeno, Miguel A., Manish Nag, Thayer S. Patterson, Andrew Shaver, and A. Jason Windawi. 2015. “The Emergence of Global Systemic Risk.” *Annual Review of Sociology* 41:65–85.

Corwin, Justin. 2002. “AI Boxing,” *SL4*, July 20. http://www.sl4.org/archive/0207/4935.html.

Future of Life Institute. 2015. “Autonomous Weapons: An Open Letter from AI & Robotics Researchers.” July 28. http://futureoflife.org/AI/open_letter_autonomous_weapons#.

Department of Defense Directive 3000.09 § 4(a)(1)(c) (Nov. 21, 2012).

Forden, Geoffrey, Pavel Podvig, and Theodore A. Postol. 2000. “False Alarm, Nuclear Danger.” *IEEE Spectrum* 37(3):31–39.

Funkhouser, Kevin. 2013. “Paving the Road Ahead: Autonomous Vehicles, Products Liability, and the Need for a New Approach.” *Utah Law Review* 2013(1):437–462.

Goertzel, Ben. 2016. “Infusing Advanced AGIs with Human-Like Value Systems: Two Theses.” *Journal of Evolution and Technology* 26(1):50–72

Good, Irving John. 1965. “Speculations Concerning the First Ultraintelligent Machine.” In *Advances in Computers*, edited by Franz L. Alt and Morris Rubinoff, 6:31–88. New York: Academic Press.

Gurney, Jeffrey K. 2013. “Sue My Car Not Me: Products Liability and Accidents Involving Autonomous Vehicles.” *University of Illinois Journal of Law, Technology & Policy* 2013(2):247–277.

Hammond, Daniel N. 2015. “Autonomous Weapons and the Problem of State Accountability.” *Chicago Journal of International Law* 15:652, 669–70.

Hernandez, Daniela. 2015. “The Google Photos ‘Gorilla’ Fail Won’t Be the Last Time AIs Offend Us.” *Fusion*, July 2. http://fusion.net/story/160196/the-google-photos-gorilla-fail-wont-be-the-last-time-ais-offend-us.

Huang, Samuel H., and Hao Xing. 2002. “Extract Intelligible and Concise Fuzzy Rules from Neural Networks.” *Fuzzy Sets and Systems* 132(2):233–43.

Hubbard, F. Patrick. 2011. “‘Do Androids Dream?’: Personhood and Intelligent Artifacts.” *Temple Law Review* 83:405–41.

Kelley, Richard, Enrique Schaerer, Micaela Gomez, and Monica Nicolescu. 2010. “Liability in Robotics: An International Perspective on Robots as Animals.” *Advanced Robotics* 24:1861–1871.

Majot, Andrew M., and Roman V. Yampolskiy. 2014. “AI Safety Engineering Through Introduction of Self-Reference into Felicific Calculus via Artificial Pain and Pleasure.” *IEEE International Symposium on Ethics in Science, Technology and Engineering* 3. doi:10.1109/ETHICS.2014.6893398.

Majot, Andrew M., and Roman V. Yampolskiy. 2015. “Global Catastrophic Risk and Security Implications of Quantum Computers.” *Futures* 72:17–26. doi:10.1016/j.futures.2015.02.006.

Marchant, Gary E., and Rachel A. Lindor. 2012. “The Coming Collision Between Autonomous Vehicles and the Liability System." *Santa Clara Law Review* 52(4):1321–1340.

McFarland, Matt. 2015. “Amazon Details Its Plan for How Drones Can Fly Safely Over U.S. Skies.” *Washington Post*, July 28. https://www.washingtonpost.com/news/innovations/wp/2015/07/28/amazon-details-its-plan-for-how-drones-can-fly-safely-over-u-s-skies/.

McMorran, Chris. 2013. “International War Crimes Tribunals.” *Beyond Intractability*. http://www.beyondintractability.org/essay/int-war-crime-tribunals.
Mracek v. Bryn Mawr Hospital, 363 Fed. Appx. 925 (3d Cir. Pa. 2010).
Müller, Vincent C. 2015. *Risks of Artificial Intelligence*. London: CRC Press.
Naughton, Keith. 2015. “Humans Are Slamming into Driverless Cars and Exposing a Key Flaw.” *Bloomberg Technology*, December 17. http://www.bloomberg.com/news/articles/2015-12-18/humans-are-slamming-into-driverless-cars-and-exposing-a-key-flaw.
O’Neill, John. 2001. “Representing People, Representing Nature, Representing the World.” *Environment and Planning C: Government and Policy* 19(4):483–500. doi:10.1068/c12s.
Page, Talbot. 2003. “Balancing Efficiency and Equity in Long-Run Decision-Making.” *International Journal of Sustainable Development* 6.
Posner, Richard. 2004. *Catastrophe: Risk and Response*. Oxford: Oxford University Press.
Robock, Alan. 2008. “20 Reasons Why Geoengineering May Be a Bad Idea.” *Bulletin of the Atomic Scientists* 64(2):14–18.
Sotala, Kaj, and Roman V. Yampolskiy. 2015. “Responses to Catastrophic AGI Risk: A Survey.” *Physica Scripta* 90(1).
Sparrow, Robert. 2007. “Killer Robots.” *Journal of Applied Philosophy* 24(1):62–77.
Sunstein, Cass R. 2006. “Irreversible and Catastrophic.” *Cornell Law Review* 91:841–97.
Tonn, Bruce Edward. 1996. “A Design for Future-Oriented Government.” *Futures* 28(5):413–31.
Vladeck, David C. 2014. “Machines without Principals: Liability Rules and Artificial Intelligence.” *Washington Law Review* 89:117–150.
Weaver, John Frank. 2014. *Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws*. Westport: Praeger.
Wiener, Jonathan B. 2002. “Precaution in a Multirisk World.” In *Human and Ecological Risk Assessment: Theory and Practice*, edited by Dennis J. Paustenbach. New York: Wiley.
Wilson, Grant. 2013. “Minimizing global Catastrophic and Existential Risks from Emerging Technologies Through International Law. *Virginia Environmental Law Journal* 31:307–364.
Wolfe, Matthew W. 2008. “The Shadows of Future Generations.” *Duke Law Journal* 57:1897–1932.
Wu, Stephen S. 2016. “Product Liability Issues in the U.S. and Associated Risk Management”. In *Autonomous Driving*, edited by Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Hermann Winner, 553–569. Berlin, Springer.
Yudkowsky, Eliezer S. 2002. “The AI-Box Experiment.” *Eliezer S. Yudkowsky*. http://www.yudkowsky.net/singularity/aibox/.

[1] The patient lost the ensuing court case, 363 F. App'x. 925, 925 (3d Cir. 2010).
[2] For further discussion of the Volkswagen case, see Baum and White (2015).
[3] Naughton, Keith. 2015. "Humans Are Slamming into Driverless Cars and Exposing a Key Flaw." *Bloomberg Technology*, December 17. http://www.bloomberg.com/news/articles/2015-12-18/humans-are-slamming-into-driverless-cars-and-exposing-a-key-flaw.
[4] The Photos story differs from the dog bite story in that Google also designed and built Photos, whereas dog owners do not design and build their dogs. But the key point remains that dogs and computer systems can both cause surprising harms.
[5] How to represent the opinions and interests of robots is a familiar question, as it also arises in the context of foreigners, future generations, non-human animals, and nature (e.g., O'Neill 2001; Wolfe 2008; Baum 2009).
[6] The legal status of these non-human entities is discussed in relation to humans and robots in Hubbard (2011).
[7] A similar argument could also apply in the event of humanity coming into contact with extraterrestrials (Baum 2010).
[8] For convenience, and without consequence to the argument, here we use consciousness and self-consciousness interchangeably. We also set aside harms to the robot that are ultimately experienced by humans, such as when the harms qualify as property damage.
[9] This need not imply that punishment is for the pleasure of members of society, as can be the case in, for example, public executions; it could instead be for their sense of justice.
[10] For related nuclear war scenarios, see Forden et al. (2010); Barrett et al. (2013).
[11] For general discussion of global systemic failure, see Centeno et al. (2015). For an example in the context of robotics, in which many self-driving cars fail simultaneously, see Baum and White (2015).
[12] On the potential for geoengineering catastrophe, see Robock (2008); Baum et al. (2013).
[13] On the possibility of global totalitarianism and the enabling role of certain technologies, see Caplan (2008); Majot and Yampolskiy (2015).
[14] Given present technology.
[15] For related brief discussion of potential approaches to AI research risk review boards and their limitations, see Barrett and Baum (2016) and Sotala and Yampolskiy (2015).