The Liability Problem for Autonomous Artificial Agents

Asaro, Peter M.

School of Media Studies, The New School

Center for Information Technology Policy, Princeton University

Center for Internet and Society, Stanford Law School

asaro@newschool.edu

Abstract

This paper describes and frames a central ethical issue–the liability problem–facing the regulation of artificial computational agents, including artificial intelligence (AI) and robotic systems, as they become increasingly autonomous, and supersede current capabilities. While it frames the issue in legal terms of liability and culpability, these terms are deeply imbued and interconnected with their ethical and moral correlate–responsibility. In order for society to benefit from advances in AI technology, it will be necessary to develop regulatory policies which manage the risk and liability of deploying systems with increasingly autonomous capabilities. However, current approaches to liability have difficulties when it comes to dealing with autonomous artificial agents because their behavior may be unpredictable to those who create and deploy them, and they will not be proper legal or moral agents. This problem is the motivation for a research project that will explore the fundamental concepts of autonomy, agency and liability; clarify the different varieties of agency that artificial systems might realize, including causal, legal and moral; and the illuminate the relationships between these. The paper will frame the problem of liability in autonomous agents, sketch out its relation to fundamental concepts in human legal and moral agency–including autonomy, agency, causation, intention, responsibility and culpability–and their applicability or inapplicability to autonomous artificial agents.

Introduction [1]

There is a growing sense of concern over the development of increasingly autonomous non-human agents–in the public and the media, as well as among policy makers and researchers. Such concerns are not unprecedented, yet there is something different about this next wave of technological innovation and change. While the impacts of the adoption of any technology are in some sense uncertain, and result in many and various unintended consequences, there seems to be something particularly unsettling and deeply uncertain about increasingly autonomous technologies. I believe that this sense of concern stems from the recognition that autonomous systems will not only be unpredictable in terms of their unintended actions and general effects on society, but that they may also be *out of control*, in the sense that these effects will occur beyond the scope of human responsibility.

Previous technological innovations have not raised this concern in quite the same way, or to the same degree, because it was possible to rely upon human morality, as well as law, to regulate the use of new technologies.[2] While there were recognized risks to using technologies like guns, cars, airplanes, steam engines, and many more, those risks were circumscribed by limits on their immediate effects and the proximate responsibility of individual humans and human institutions for those effects. In most cases, these risks were also limited in their complexity and longevity, in the sense that those effects could be anticipated, managed, and reduced over time.[3] As long as there is a human hand on the controls, there is a margin of assurance that the technology cannot go too far out of control before it is reigned in by human agency or regulatory policy. Thus, human control presents itself as a

1 Research for this paper was supported by a grant from the Future of Life Institute.

2 There are some notable exceptions. There have been debates over the morality of research and development of technologies with potentially catastrophic results or applications (and sometimes the dissemination of basic research in these areas) particularly the weaponization of nuclear physics and biological and chemical agents. There have been similar debates over the risks of basic research itself in the case of genetic engineering and nanotechnology, where the experiments themselves might go "out of control" and cause irrevocable or catastrophic harm. Many of these concerns are related to the potential causal "chain reactions" of positive feedback systems and exponential growth.

3 Notable exceptions here are certain tragedies of the commons and the problem of many hands (Nissenbaum 1996), in which proximate responsibility is small and requires many participants or instances for the harm to be realized. Environmental degradation is a good example where the participation of many separate responsible agents leads to a diffusion of perceived responsibility. Such harms are difficult to regulate largely because it is more difficult to establish the causal link to the responsible agents and rule out intervening or contributory causes.

"kill switch" to technology going completely out of control, while human responsibility further acts as "policy lever" for policy to enact regulation over material technologies.

Non-human agents, including robotics and software agents, and especially those using advanced artificial intelligence (AI), are becoming increasingly autonomous in terms of the complexity of tasks they can perform, their potential casual impacts on the world that are unmitigated by human agents, and the diminishing ability of human agents to understand, predict or control how they operate. This increasing autonomy of non-human agents seriously challenges human control, and thus challenges both the "kill switch" and the "policy lever" functions of human control. Some have argued that one factor or other is more troubling, or that one challenge or the other is more dire. But, I believe that the growing concern over increasingly autonomous agents stems from the combination of these factors and the challenges they impart, and they must be addressed comprehensively and systematically. Central to addressing these challenges will be finding a way to manage the liability problem–how to continue to hold people legally liable for increasingly autonomous agents. This is closely related to the responsibility problem–how to ensure that human agents take moral responsibility for the technologies they make and use as these become increasingly autonomous.

How should we regulate artificial computational agents, including AI and robotic systems, as they become increasingly autonomous, and supersede current capabilities? In order to take a systematic approach to this question, we must integrate an analysis of fundamental concepts of autonomy and agency, with a review of current processes of systems design and development, and chart a course for future regulatory policy. This will include investigating 1) the conceptual foundations of autonomy and agency; 2) survey the existing legal and moral theories of liability and responsibility and their applicability to artificial agents; and 3) explore how implementing an enhanced theory of liability and agency for autonomous artificial agents could shape regulatory policy going forward. In this paper I will focus on sketching out 1) the conceptual foundations of autonomy and agency, through an examination of the liability problem.

One of the central problems facing the development of autonomous artificial agents as technological capabilities continue to increase, is the uncertain status of liability for the effects caused by artificial agents–the liability problem. Resolving the liability problem will require untangling a set of theoretical and philosophical issues surrounding causation, intention, agency, responsibility, culpability and compensation. The primary objective of the proposed research is to address this problem through a conceptual analysis that is informed by existing legal and moral theories, and grounded in technological capabilities of current and potential artificial agents.

The Liability Problem for Autonomous Artificial Agents

In order for society to enjoy many of the benefits of advanced AI and robotics, it will be necessary to be able to deal with situations that arise in which autonomous artificial agents violate laws or cause harm. If we want to allow AIs and robots to roam the internet and the physical world and take actions that are unsupervised by humans–as might be necessary for, *e.g.* personal shopping assistants, self-driving cars, and host of other applications–we must be able to manage the liability for the harms they might cause to individuals and property. Already, there have been automated shopping software bots that have purchased illegal items on the DarkNet, albeit as an art project (Kasperkevic 2015). There have also been serious questions raised as to who could be held accountable for the deaths caused by autonomous weapons in war (Human Rights Watch 2015).

Traditional approaches to handling liability are inadequate for dealing with autonomous artificial agents due to a combination of two factors–unpredictability, and causal agency without legal agency.

First, unlike traditional engineering and design, the actual functioning of an autonomous artificial agent is not necessarily predictable in the same way as most engineered systems. Some artificial agents may be unpredictable in principle, and many will be unpredictable in practice. Predictability is critical to current legal approaches to liability. In traditional product liability, the manufacturer is responsible for the product working as designed, and foreseeing likely problems or harms it may cause. Determining what is "foreseeable" often falls upon courts to decide, but the legal standards used are whether the manufacturer had knowledge of the potential problem, or whether a reasonable person should have foreseen it, or whether there is an industry standard of practice that would have revealed it. While there is a degree of unpredictability in the performance of any engineered product, due to failures or unforeseen circumstances of use, there are shared expectations regarding its performance, testing for the limits of that performance and likelihood of failure, and management of foreseeable risks.

In the case of advanced AI, a system that learns from environmental data may act in ways that its designers have no feasible way to foresee (Asaro 2008). In a limited sense, this is already the case with current machine learning techniques, which use large sets of training data and produce novel problem solutions. Currently, it is possible to analyze and test a learned function and determine its behavior, as with traditional engineering. But when AI systems are allowed to continue modifying their functions and learn after they are deployed, their behavior will become dependent on novel input data, which designers and users cannot predict or control. As a result, the behavior of the learned functions will, to various degrees, also be unpredictable. A truly robust AI program capable of open-ended learning could learn functions that

its manufacturer could not foresee, perhaps in principle. Insofar as autonomous artificial agents utilize learning and open-ended learning, their behavior will also be unpredictable. Unpredictability by itself is not an insurmountable problem for liability, insofar as the agents who introduce that unpredictability could be themselves held liable, or the risks from unpredictability could be managed.

However, the second factor challenging traditional approaches to liability is that autonomous artificial agents may "act" on their own, yet are not accountable or liable in a legal sense. With most engineered products, the predictability is bounded by the actions of other agents–consumers, users, service technicians, *etc.*–and how they maintain and use a given product.[4] In those cases, there is a clear legal agent or user who may have used the product inappropriately and is thus liable for the consequences of the use in that instance (at least partially, if not completely).

Autonomous artificial agents can act in the world independently of their designers or operators. This makes it difficult to identify the user or operator, who would normally be liable. In the case of learning systems, the causal influence resulting in the unpredictability of a system stems from the datasets used for learning, not a legally responsible agent (unless someone deliberately seeks to influence the datasets and learning process). These cases are somewhat analogous to the way that parents may be held liable for the actions of small children, but they are not usually held liable for the actions of their adult children. As adults, the children have learned enough about the world to become their own legal agents. Something similar applies in cases of employer-employee liability. Usually the employer is responsible for any damages their employee causes in the course of doing their job (such as damage from a delivery truck hitting a parked car while making routine deliveries), but if the employee is off on a frolic (and took the delivery truck to visit a friend), then the employer is not responsible for the damages and the employee is (Asaro 2011).

Approaches to the Problem

Taken together these two factors result in a number of problems with applying traditional theories of legal liability to autonomous artificial agents. There are two basic liability frameworks in the law, criminal and civil (primarily tort) liability. In both domains, it is difficult to hold the artificial agent legally liable for its actions, as they are not legal persons, and treating them as legally fictitious persons (like corporations) does not really solve the problem (Asaro 2011). It is possible in some criminal situations to hold the people who operate the artificial systems liable, provided you can show intent to commit a crime, or foreseeable risk of a harm rising to the level of criminal negligence. To the extent that artificial agents become increasingly complex, those who build or deploy advanced AIs and robotics will not necessarily have intent or foresight of the actions those systems may take. This is especially true for systems that can substantially change their operations through advanced learning techniques, or those future systems that might even become genuinely autonomous in generating their own goals and purposes, or even intentions.

At some point in the evolution of autonomous artificial agents, they might become legal and moral agents, and society will be faced with the question of whether to grant them some or all of the legal rights bestowed on persons or corporations. At that point, some or all of current liability law might apply to those AIs and robots that qualify as legal persons, though it might not be clear what the exact boundaries of a particular entity might be, or how to punish it (Asaro 2011). For example, whether a particular instantiation of a program is the legal subject, or all copies of a program are part of the same legal subject (as copyright law might suggest). It will also not be clear how to appropriately punish them, or otherwise correct their future actions and provide retributive compensation to those that have been harmed. Despite these challenges, being able to treat autonomous artificial agents as legal persons is likely to be the easier problem.

In the near term, the hard liability problem for autonomous artificial agents will lie in devising a system of liability that promotes beneficial innovation while offering just and adequate compensation to those harmed. Justice requires that those who will be held liable should be able to understand the scope and extent of the risks and liability they are assuming in deploying an autonomous artificial agent, and have some means of managing that risk through controls over the system. Adequate compensation means that there needs to be sufficient means to compensate those harmed, monetary or otherwise–whether this entails holding large corporations accountable, or providing proper risk-pooling insurance. For insurance to work, of course, it will be necessary for actuaries to be able to assess the risks posed by various artificial agents. Assessing those risks will require both expertise, and probably means for imposing or assuring predictability in the systems themselves. And even then, there may be catastrophic risks involved, which may not be manageable under these frameworks. Ultimately there will be questions that society must address as to the perceived benefits and risks of such technologies, and who in society should receive those benefits and should shoulder the burden of those risks.

There are a few liability policy options already in use in tort law, such as joint and several liability, strict liability, and risk-pooling insurance that might be deployed. But there are reasons to doubt that these will scale adequately

4 For example, a hammer is designed to drive nails, but can also be used as a weapon–with the liability falling on the person who uses it as a weapon. Similarly, the car owner is responsible for changing the brake pads of a car when they are worn, and the mechanic is responsible for doing it properly, while the manufacturer's liability is limited to properly informing car owners of the service needs for the car.

or address all the problems posed by autonomous artificial agents.

Existing forms of joint and several liability permit those harmed to seek monetary damages from the deepest pockets among those parties sharing some portion of the liability. This works well where there is a large corporation or government that is able to pay damages. While the amount they pay is disproportionate to their contribution to the damages, those harmed are more likely to be adequately compensated for their damages. One result of this is that those manufacturers likely to bear the liability burden would seek to limit the ability of consumers and users to modify, adapt or customize their advanced AI and robotics products in order to retain greater control over how they are used. This would stifle innovation coming from the hacking, open-source and DIY communities. The hacking, open-source and DIY communities, while a powerful source of innovation, will have limited means of compensating those who might be harmed from their products—not just those who chose to use them.

Strict liability goes further and designates or stipulates a party, usually the manufacturer or owner, as strictly liable for any damages caused. This model of liability applies to such things as the keeping of wild animals. It is expected that tigers will harm people if they get free, so as the keeper of a tiger you are strictly liable for any and all damages the tiger may cause. We apply normal property liability to domesticated animals, however, which we expect to not harm people under normal circumstances. It has been suggested that we could apply this to robotics (Schaerer, Kelley, and Nicolescu 2009), and perhaps designate certain advanced AIs as essentially wild animals, and others as domesticated. But then the question arises: How does one determine whether an advanced AI or robot is appropriately domesticated, and stays that way? A designated liable person may not know the extent to which the autonomous system is capable of changing itself once it is activated. More problematic, however, is that a system of strict liability might result in the slow adoption of beneficial AI technologies, as those who are strictly liable would have a large and uncertain risk, and be less likely to produce or use the technology, or soon go out of business.

There are other proposals that aim to design mechanisms of accountability into complex socio-technical systems (van den Hoven, Robichaud, and Santoni de Sio 2015). But such approaches, much like strict liability, do not really address the fundamental issue of autonomous agency–how to define it or to regulate it. They merely try to manage it in an *ad hoc* effort to maintain traditional concepts rather than find a simpler and more powerful solution.

We thus have a few general options when it comes to regulating autonomous artificial agents. We could preclude or prohibit autonomous artificial agents because of their risks and uncertainties. We might call this the precautionary approach. To the extent that it works, it would also impede the development and deployment of many beneficial advanced AIs and robots because of their problematic autonomy. We could permit the development and deployment of autonomous artificial agents, and accept the risks and costs at a social level, without developing a better framework for regulating autonomy. We might call this the permissive approach.[5] This would allow many beneficial applications of advanced AI and robots, but also many harmful ones, including many harms for which no one might be liable and those harmed would not be compensated. As a secondary effect, there would likely be a general backlash against advanced AI and robotics as the technology comes to be seen as harmful and offering few options for restitution of those harms.

Or, we could pursue one of the heavy-handed liability schemes, such as strict liability, that would regulate the industry to some extent, but also limit innovation to those areas where there are sufficient profits to motivate large capital companies to enter the market and accept the risks. We might see this work in applications such as financial trading algorithms, self-driving cars and medical-care robots, but many other beneficial applications of AI may not have such clear profit margins. Even in clearly profitable areas, strict liability is likely to stifle innovation for a range of beneficial applications of autonomous artificial agents.

Alternatively, we could seek a better solution to the liability problem than current models afford. We could attempt to reconceptualize how we think about agency, causality, liability responsibility, culpability and autonomy for the new age of artificial autonomous agents. While it is not yet clear how this might work, a clear framing of the problem, as this paper has presented, is the first step in that larger project. The next steps involve clarifying the distinctions between the purposes of algorithms and machines from the purposes of the persons who use them (Asaro forthcoming), and distinguishing different types of agency accordingly.

References

Asaro, Peter (forthcoming) "Roberto Cordeschi on Cybernetics and Autonomous Weapons: Reflections and Responses," *Paradigmi: Rivista di critica filosofica*, 2016.

Asaro, Peter (2008) "From Mechanisms of Adaptation to Intelligence Amplifiers: The Philosophy of W. Ross Ashby," in Michael Wheeler, Philip Husbands and Owen Holland (eds.) *The Mechanical Mind in History*, Cambridge, MA: MIT Press, pp. 149-184.

Asaro, Peter (2011) "A Body to Kick, But Still No Soul to Damn: Legal Perspectives on Robotics," in Patrick Lin, Keith Abney, and George Bekey (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics.* Cambridge, MA: MIT Press, pp. 169-186.

Chopra, Samir and Laurence F. White (2011) *A Legal Theory for Autonomous Artificial Agents*, University of Michigan Press.

5 This is the approach taken by Chopra and White (2011).

Human Rights Watch (2015) “Mind the Gap: The Lack of Accountability for Killer Robots,” HRW Report, April 9, 2015, 38pp. Downloaded from: http://www.hrw.org/node/133918

Jeroen van den Hoven, Jeron, Phil Robichaud, and FillipoSantoni de Sio (2015) “Why the Future Needs us Today: Moral Responsibility and Engineering Autonomous Weapon Systems,” Address to the United Nations Convention on Certain Conventional Weapons Expert Meeting on Lethal Autonomous Weapons. Geneva, Switzerland, April 16, 2015. Downloaded from: http://www.unog.ch/80256EDD006B8954/(httpAssets)/89116E298AE593C2C1257E2A00413D61/$file/2015_LAWS_MX_VanDenHoven.pdf

Kasperkevic, Jana (2015) “Swiss police release robot that bought ecstasy online,” *The Guardian*, April 22, 2015. Downloaded from: http://www.theguardian.com/world/2015/apr/22/swiss-police-release-robot-random-darknet-shopper-ecstasy-deep-web

Nissenbaum, Helen (1996) Accountability in a Computerized Society,” *Science and Engineering Ethics,* March 1996, Volume 2, Issue 1, pp 25-42. Downloaded from: http://www.nyu.edu/projects/nissenbaum/papers/accountability.pdf

Schaerer, Enrique, Richard Kelley, and Monica Nicolescu (2009) “Robots as animals: A framework for liability and responsibility in human-robot interactions,” *Proceedings of RO-MAN 2009: The 18th IEEE International Symposium on Robot and Human Interactive Communication*, Toyama, Japan, Sept. 27–Oct. 2, 72–77.