THE FUTURE OF THE GLOBAL ORDER COLLOQUIUM > FALL 2019

Responsibility and Autonomous Technologies:

Is there a problem?

By Claire Finkelstein, *Algernon Professor of Law and Professor of Philosophy, University of Pennsylvania Carey Law School; Founder, Center for Ethics and the Rule of Law (CERL)*

There is a common worry expressed in popular as well as scholarly writings about autonomous technologies, namely that they render the question of responsibility for wrong or harmful actions ambiguous. The fear is that if self-driving cars or autonomous weapons systems are themselves making the decisions about what route to take or whom to target, and if there is truly no "human in the loop" in these cases, there will be no one to hold responsible for injury that occurs. This being the case, human beings will be able to exploit autonomous technologies in order to commit all manner of egregious acts but avoid responsibility for the harm they inflict. At the very least, we will find ourselves in a confusing situation of multiple and overlapping domains of responsibility, with no one whose clear duty it is to prevent harms inflicted by autonomous systems. The purpose of this paper is to assess whether these concerns are warranted and to ask the question whether developments in autonomous technologies do pose a fundamental threat to the nature of our traditional responsibility judgments. My argument will be that in the domain of classic tort or criminal law judgments for harm, there is very little difference between a wrongful harm inflicted directly by a human being and a harm inflicted by an autonomous system. Both in law and in morals, our traditional judgments can be fairly straightforwardly applied to autonomous processes. In the domain of cyber, however, there are fundamental changes taking place, facilitated by autonomous processes, where responsibility becomes significantly more diffuse. In that domain, responsibility will have to be prospective, and duties imposed despite lack of clear causality, given the present rather limited state of our knowledge.

I.

In much of the literature about AI, there is a sense that if computers are deciding what to do on their own, then no one could be responsible for the harm inflicted by autonomous technology. The most obvious place in which the worry about autonomy and responsibility comes up is with malfunctioning self-driving vehicles or improperly functioning technologies such as autopilots. Recently, a technological miscalculation caused an accident with a bus for one of Google's self-driving cars when it changed lanes and crashed in the side of a vehicle. In another accident, a self-driving Tesla failed because it was blinded by the sun, and the driver who should have intervened was busy watching a Harry Potter movie. Normally, of course, the driver of the vehicle would be responsible for a crash, but what if there is no driver?

A more significant real-world example is playing out currently with regard to a feature of the autopilot on the Boeing 737 Supermax, the failure of which has caused two catastrophic crashes. Investigators have now traced the cause of both crashes to the Maneuvering Characteristics Augmentation System (MCAS), which is unique to the Supermax. The system automatically brings the nose of the plane down, but only under a very narrow set of circumstances. There was a ready solution to this problem that required a manual maneuver on the part of the pilots, but without proper training and a clear identification of the problem in advance, pilots were at a loss when the system malfunctioned. The situation, however, is deeply complex, since causally the fault in some sense lies at the interface of human control and malfunctioning autonomous equipment, and the two cannot be fully disentangled. Indeed, what counts as malfunction for any piece of equipment depends on what the baseline human capacities are assumed to be: in the case of well-trained, experienced pilots who were aware of the risk posed by the autopilot system, that system did not ultimately pose a significant problem. But the pilots of the two planes that crashed did not have the requisite training, and against that backdrop, the defect in the autopilot system was unmanageable.

While there are causal complexities in situations involving automated or autonomous technologies, in fact automaticity or autonomy adds nothing new here, at least nothing that can't be addressed using a variety of traditional legal principles. One such principle is that of vicarious liability: if you keep a wild tiger in your back yard, and it escapes and kills the toddler next door, you cannot be heard to assert that the tiger is an "autonomous" creature that selected its own target and that responsibility is murky because there is no human in the loop. You are as responsible as if you had intentionally left an ice slick on your sidewalk intending passersby to slip, or left your loaded AK-47 on your front lawn for anyone to pick up and use. Bartenders are liable for the acts of their visibly inebriated customers if they serve them drinks while in that condition, parents are liable for the acts of their young children, and employers are frequently liable for the acts of their employees. Thus there will be no difficulty holding Google responsible for the mistakes of their self-driving cars, and Boeing liable for the unexpected movements of its autopilot technology, premised on reasonable expectations about pilot response.

Moreover, the case for liability in the examples of autonomous technologies is even stronger than in the case of wild animals. Where autonomous technologies are concerned, we have designed the "animals" and thus are responsible for the autonomous processes in at least two senses: 1. We have direct "design" responsibility based on the fact that we have created the system that is causing injury, and 2. We have responsibility for failing to take adequate precautions against a foreseeable, indeed in some cases foreseen, harm we could have prevented. Thus self-driving cars or autopilot technologies do not appear fundamentally to confound the basic responsibility judgments we are able to draw. We have enough tools in our toolkit to handle the vast majority of cases that arise. Where we lack such tools, it is because they are hard cases, and traditional legal principles would have had difficulty with even the most quotidian of instances, irrespective of the technology involved.

II.

So far so good. But in another domain of our lives that intersects with autonomous technologies, matters may not be so simple. Social media is now populated by numerous bots that, combined with human various human purposes, assist in replicating and distorting content across the globe. If you give a televised interview, your words may be transformed and disseminated, your face may be copied and your image reproduced in multifarious places you did not intend or to which you did not give your permission, autonomously functioning programs may translate your words (and distort them) into Russian, Chinese, Korean, and so on, and you may find yourself cited for things you never knew you said. If you are using voice activated technologies, the selections you make and the words you use may be recorded, and it is not difficult to imagine that voices could be transformed and altered to seemingly place words in someone's mouth, as is starting to occur in abundance. These "deep fakes" have the potential to drastically reduce the reliability of internet content, with the potential for particular damage to political and other public figures.

There is already a significant risk to our personal privacy and security. Personal information collected by voice technology or ordinary internet data collection has the potential to expose all your likes or dislikes, your search habits, the sites you regularly visit, and the news you read. All this can be identified, recorded, shared with third parties and disseminated across the internet. Indeed, from automatic tracking of your on-line behavior, it appears that the internet knows whether you are depressed before you do, just by tracking changes in your social media posts and the types of sites that you explore. The vulnerability of individual internet users through on-line exposure, the collection of information, trolling, and confusion about the source of one's own or other's political positions, and the potential for deep fakes with regard to any on-line content poses perhaps the greatest challenge to freedom of expression and participation in public discourse in our history.

Consider the case of a chatterbot called "Tay," which was released by Microsoft via Twitter in March of 2016. Tay was designed to mimic the tweeting patterns of a 19-year-old American girl, and to learn from interactions with other users. Within a few hours of its launch the bot began to post inflammatory and offensive tweets through its Twitter account, forcing Microsoft to shut the service down only 16 hours after it launched. The bot had "learned" through racist and offensive contacts with users on Twitter. Microsoft asserts that the bot was "attacked" by users intentionally seeking to teach it offensive phrases. Whether intentional or not, however, the capacity to absorb the offensive content floating around on the internet runs the risk of this sort of technology "learning" the worst of what social media has to offer, and a risk of its then influencing users in turn.

The stakes are high—higher than this particular example suggests. While the internet has been a place of increased political engagement—a good thing for reinforcing democratic norms, we also know now that political content shared on social media is seriously untrustworthy, perhaps more untrustworthy than any other content we counter in cyberspace. There is increasing evidence that chatbots released on Twitter and other social media increase political polarization and thereby contribute to the instability of democratic governance. Thousands of bots are enlisted to push out and respond to political content, and evidence suggests that Russia and possibly other foreign governments have seen themselves as advantaged by increasing dissention among the United States voting population and sewing dissent.

Moreover, the presence of bots in our online conversations distorts our interactions on social media and gives the illusion that greater political support may exist for some positions or causes than truly does. Indeed, propaganda bots may be influencing your thoughts in ways you are not aware of. While we think we are interacting with genuine, even thoughtful interlocutors in Twitter, Facebook or other platforms, your conversational partner may not in fact be human at all, and it may be planting ideas in your head that are largely put there by a foreign power. We, in turn, are responding by developing further reflections to augment or combat these inanimate conversationalists, but our responses to internet chatter may themselves reflect what we are being fed. Although the less educated we are, the more vulnerable we are to internet trolls, the educated elite has at least been deeply affected by foreign source manipulations. Ultimately instead of the bots echoing us or reflecting public opinion, we may come to be reflecting them.

Judgments of responsibility for this reversal of our political dialogue and the increasing automaticity of our political reflection are more complex that the cases of personal injury at the hands of automatic and autonomous navigational technologies. Rather than the model of responsibility for the acts of a wild animal or a child, these technologies engage in wrongdoing that is highly diffuse in origin. If a chatterbot is simply absorbing and reflecting racist and other offensive chatter that exists in our larger society, is the "design responsibility" we have for the behavior of those systems adequate to explain what is going on? Is it an accurate and fair representation of our principles of responsibility? We impose liability on parents for the acts of their children, but once the children grow up and have "learned," we are no longer responsible for what they do, especially if what they have learned is prevalent in the society in which they live.

The situation is more similar to the diffuse causality involved in polluting a river when multiple industrial centers are contributing to that pollution. It may be that no single factory dumping into a common waterway is responsible for the contaminated state of the river. But taken together, a threshold is passed above which the waterway cannot repair itself. The process is highly interdependent: it depends on the behavior of other actors over which any single manufacturer lacks control. Foreseeability and so-called proximate cause analysis are impossible to apply here, for unlike the case of Boeing and its autopilot, the separate behavior of all the different polluters is only toxic once they reach a certain level, a level that requires no coordination among the actors involved. In short, there is no "but for" causation in the case of the polluters, and thus arguably no way to pin responsibility on any individual actor.

Is the damage we have allowed to occur, and continue to allow, to our political thought and dialogue like the polluted river? Should Microsoft, Google, Facebook, etc. be held liable for that which is only partially their fault, and where their intervention is a necessary but not sufficient condition? As the cyber domain grows in power and complexity, questions of responsibility will require deeper and more careful analysis. We shall have to determine, in very short order, whether our legal and moral concepts are up to the task.

Disclaimer: *This article was drafted for the 2019 Global Order Colloquium at Perry World House, the University of Pennsylvania's global affairs hub, and made possible (in part) by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.*

Claire Finkelstein is the Algernon Professor of Law and Professor of Philosophy at the University of Pennsylvania Carey Law School and founded the Center for Ethics and the Rule of Law (CERL), a non-partisan interdisciplinary institute that seeks to promote the rule of law in national security, warfare, and democratic governance.