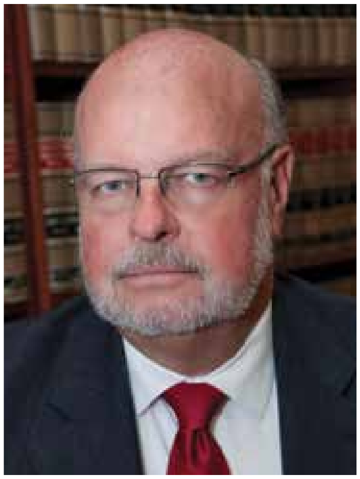
# Artificial Intelligence, Criminal Liability and the Trolley Problem

By Robert M. Sanger



Robert M. Sanger

Most people today are familiar with the "trolley problem." It was first stated by philosopher Phillipa Foot in 1967.[1] It was made popular by Harvard Professor Michael Sandel in his introductory lectures at Harvard University which were broadcast by BBC television and by the University online in 2016.[2] The trolley problem was even featured in several television shows, most famously in the NBC series *The Good Place*, a show that incorporated a course in ethics with an intellectually robust sitcom.[3] The trolley problem has been the subject of popular books as well.[4]

Most people are also familiar with the litigation related to machine generated decisions, particularly in self-driving vehicles. Tesla's use of artificial intelligence (AI) was the subject of a civil trial in Los Angeles which resulted in a defense verdict.[5] One of the issues regarding self-driving vehicles is the extent to which there is corporate liability for moral choices that are programmed into the vehicle or the extent to which it was operator error. However, Tesla and others have been largely successful in shifting liability away from the corporate defendants to the individual operator. Nevertheless, as of last year, Tesla's self-driving vehicles have been reported to have been involve in over 736 crashes, including 17 fatalities.[6] Looking just at the string of collisions with first responders while the driving automation system was engaged, the National Highway Transportation Safety Administration (NHTSA) has investigated at least 11 crashes in nine states.[7]

AI programming has also come under scrutiny in the deployment of Lethal Autonomous Weapons (LAW).[8] There have been numerous stories of innocent civilians and even children being killed by AI guided drones. Looking back over a decade, AI assisted drones were used by the United States in Pakistan to effect "targeted killing" of Taliban leaders. Nevertheless, many civilians were killed, including children, in an effort to hit the targets. For instance, in the course of two "targeted" AI assisted drone strikes, the United States attempted to kill a single mid-level Taliban functionary. It was unsuccessful—the target survived to become, years later, an al-Qaida leader. However, the efforts cost the lives of 76 children and 29 adults.[9]

The "trolley problem" thought experiment can be helpful in a discussion of the moral implications of making choices and allocating responsibility between corporate defendants and the individuals in both the civil and criminal contexts. A human operator of a self-driving vehicle or a lethal autonomous drone may do an "act" that results in the death of a third party. That death, if caused by the negligent or unlawful act could cause civil or criminal liability for the individual, corporate or military actor. Civil liability has been the subject of a good deal of literature but much less so criminal.[10]

This *Criminal Justice* column will focus on one aspect of potential criminal liability as divided between corporate and individual responsibility, specifically the efforts of corporate and governmental entities to retreat from grandiose claims of the superiority of AI and, now, to publish cautious caveats claiming that these self-driving cars and lethal autonomous drones are really dependent on the control of individual operators. This may be a correction for past hubris but the consequence is that individuals may be sacrificed for the benefit of the corporation or the government. In self-driving vehicle collisions this move may decrease the civil and (rarely imposed) criminal liability on the part of the corporation, but it may enhance the potential civil and (more often imposed) criminal liability on the part of the individual. In lethal automated drone cases, the government may give plausible deniability to claims of human rights violations while possibly exposing individual operators to a war crimes prosecution.

## *The Trolley Problem*

As phrased by Phillipa Foot in her original 1967 article:

> "To make the parallel as close as possible [to another analogy] it may rather be supposed that he is the driver of a runaway tram which he can only steer from one narrow track on to another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed."[11]

This was part of an effort to unpack the idea of unintended consequences or what Foot referred to as the "double effect." It is of note that the trolley analogy was one of many scenarios that posed moral dilemmas. Each variation on the theme involved a choice between actions (or inactions) that resulted in harm whichever choice was made. While the paper discussed the double effect in abortion decisions, it turned to an intentionally humorous example of a large man[12] who gets stuck in a cave trapping his fellow spelunkers behind him while the water level starts to rise. One of his trapped colleagues happens to have a stick of dynamite and the moral choice presents itself as to whether all the trapped explorers should die or whether the man should be blown up with the dynamite.[13]

Of the many other double effect thought experiments, one involves a pilot's choice to land a plane to save all on board or to save fewer people on the ground;[14] another, is the villagers' choice to turn in one innocent person to spare destruction of the village by the gestapo;[15] a third is the *Dudley and Stevens* choice between sacrificing the cabin boy and death of the others on board.[16] All of these choices are similar to the moral challenge faced by the actor in the trolley problem.

Philosophers have debated the trolley problem endlessly and there is no convincing consensus. Modifying the underlying facts makes the analysis all the more difficult. Perhaps Michael Sandel has an answer of sorts. He concludes that fundamental philosophical quandaries, like the trolley problem, seem to call for a pat analytical resolution. Even if there is an analytic answer based on the prophylactic hypotheticals, what would happen in reality is multi valent and inherently uncertain. Therefore, according to Sandel, the best solution is to engage in informed intellectual discourse on the subject.[17]

In the case of AI, intellectual discourse is replaced by probabilistic determinations that are based on preprogramed algorithms and machine learning, along with the contemporaneous input received by its sensor system. Bottom line: in self-driven vehicles or LAW's, the life or death choices presented in the trolley problem are, de facto, decided by the machine.

### Trolley Switches, Self-Driving Vehicles and Lethal Autonomous Weapons

There are issues of potential civil and criminal liability for, for instance, choosing to kill five people rather than opting to sacrifice one. The responsibility for that choice could be assessed against the individual operator of the vehicle or, in some other analogous situation, the operator of the drone. On the other hand, if the AI programming led to this "choice," there is an argument establishing liability on the part of the individuals or entities involved in the design, engineering, manufacturing, programming, or placement of the self-driving vehicle in commerce. For the sake of this discussion we will consider the last group of individuals and entities in the aggregate as the corporate manufacturer or the government.[18] That is the issue: the subtle shifting of liability from the corporate manufacturer or government to the individual operator.

The term "self-driving vehicles," according to the Society of Automotive Engineers really is broken down into six levels from "Level 0 (no driving automation) to Level 5 (full driving automation) in the context of motor vehicles and their operation on roadways."[19]

Level 0 is a vehicle with the basic late model vehicle equipment, including limited warnings, automatic emergency braking, blind spot warnings. Level 1 includes some steering or brake/acceleration support, lane centering or cruise control. Level 2 includes all of the features that may be options in Level 1. Levels 0, 1, and 2 all require an active driver who must supervise the features and steer, brake or accelerate as needed.

Levels 3, 4, and 5 are such that no individual operator is driving when the automated features are engaged, and, only in Level 3, the driver must drive when the system requests. In Levels 4 and 5 the automated features will not require or permit driver input and may not even have pedals or a steering wheel installed. Level 4 would include driverless taxis but can only operate where specific conditions are met. Level 5 can drive under all conditions.

Corporate liability in Level 0, 1, and 2, would involve typical products liability issues, otherwise, if there is liability, the individual operator would be responsible. At these levels, if there is time and capability, and the driver could override the AI choices, the ultimate liability may be individual or joint. In litigation over a collision, there would be a conflict between the individual operator and the corporate manufacturer. Yet, where the operator is passive for the most part in Level 3, or where there essentially is no operator as in Level 4 and 5, if there is liability, it would be corporate.

Lethal Autonomous Weapons (LAW) are subject to a similar analysis. Department of Defense Directive (DoDD) 3000.09, "Autonomy in Weapon Systems," was originally published November 21, 2012 and after 10 years was just updated earlier this year on January 25, 2023. The purpose of the directive is to establish "guidelines designed to minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements."[20] While there is some human input, the Lethal Autonomous Weapons operate on AI to deliver lethal force. Just like the algorithms that are programmed into self-driving cars, programed algorithms and machine learning processes make "choices" leading to the killing of human beings.[21]

## *Placing Potential Blame on the Human Operator*

Currently, there are limited examples of Level 4 operating vehicles, such as Waymo taxi cabs operating in limited areas and under somewhat controlled circumstances. Full Level 5 vehicles are still experimental. Tesla, for instance, seemed to be claiming that their Full Self Driving vehicles were Level 3 or 4, but they have strategically backed off any such apparent claim. Tesla, Inc. issued the following statement in an SEC filing last year:

> "Currently, we offer in our vehicles certain advanced driver assist systems under our Autopilot and FSD options. Although at present the driver is ultimately responsible for controlling the vehicle, our systems provide safety and convenience functionality that relieves drivers of the most tedious and potentially dangerous aspects of road travel much like the system that airplane pilots use, when conditions permit. As with other vehicle systems, we improve these functions in our vehicles over time through over-the-air updates."[22]

Note that Tesla is downplaying its "Full Self Driving" capability (which sounds like a Level 5) by saying that, despite the name, their "FSD" vehicles are no more than a SAE Level 2.

The Department of Defense updated their DoD Directives in 2022 on Lethal Autonomous Weapons to specifically include language that imposes a standard of care on DoD personnel for development, deployment, and use: "DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities."[23] This was in part due to a statement a year earlier by General Mike Murray, head of the Army Futures Command. He told reporters that, "Where I draw the line — and this is, I think well within our current policies—[is], if you're talking about a lethal effect against another human, you have to have a human in that decision-making process."[24] At the time he said it, that was not DoD Policy when the General made the statement but now it is.[25]

Therefore, there is an acknowledgement, as evidenced by Tesla policy on FSD and the Department of Defense Directive, that a human being is accepting responsibility for the actual use of potentially lethal devices that are directed by AI. Tesla has been somewhat successful in actual litigation to divert liability from the corporate manufacturer to the "operator." The Department of Defense may have been less motivated by a concern for liability as for public relations in issuing its updated directive but, again, it places potential liability on an individual operator where, for instance, a preschool was targeted by AI in the midst of armed conflict.

## *Liability for a Choice that Causes Death*

So, back to the "trolley problem." When a device employing AI is the cause of a death, liability could be placed on the corporate or governmental entities, or on the individual operator if there is a showing that the cause of the death was negligent, a violation of law, or intentional, or grounded in strict liability. In criminal law, the individual operator could be prosecuted for misdemeanor manslaughter, or felony manslaughter with gross negligence, manslaughter while under the influence or a combination of gross negligence and under the influence.[26] If the prosecutor could meet the standards, an operator could be prosecuted for a *Watson* murder.[27] Similarly, a Lethal Autonomous Weapon that uses AI to define targets may kill civilians and violate the Geneva Convention. Individual operators could be held liable under international law.[28]

Decisions that trouble humans, such as the "Trolley Problem," are simply cold calculations for AI. AI's "choices" are based on probabilistic determinations enhanced by machine learning rather than critical thinking about what is right or wrong. However, there are people making corporate decisions, who marketed a product and who made their choice to put people at risk. The infamous Pinto case, is an example of a crass economic calculation to put a vehicle on the road that might explode. In self-driving vehicles, the choice may be more subtle—somewhere in the evolving algorithms, AI will make a choice based on probabilistic projections to kill one rather than the uncertain risk of killing five. That "choice" may be bad enough but then consider uncertainty in the real world which means that the choice may result in death where all lives might have been saved.

In defense of AI, one could argue that all decisions about the real world are rooted in uncertainty. AI may claim it makes better choices on the average than humans. And, it is a fact of society, particularly postindustrial society, that people risk being harmed by new technology. It is a fact that ordinary automobiles kill people and yet we risk going out in traffic because we have accepted that the convenience of mobility is worth the risk.

That is what might be the downfall of those arguing that self-driving vehicles are "worth the risk." They are "bright shiny objects" but hardly provide the kind of benefit that the car itself does. The benefit to a self-driving vehicle is that a person does not have to pay (as much) attention while behind the wheel. Advertisements show people singing and clapping their hands. Is that worth killing a number of people to be able to do that?

So, self-driving vehicles or lethal autonomous drones can be designed or marketed in such a way that the corporate of government parties could be held liable for a civil judgment or, conceptually, for a criminal judgment. It is the attempt to avoid big judgments in automobile collisions or, in the case of LAW's, to avoid international reproval. But that comes with a price to the individual operators. The individual operators generally are overpowered by corporate or government lawyers and this imbalance of power will lead to more findings of individual responsibility, perhaps where it is not warranted. In turn that leads to more criminal prosecutions of vulnerable individuals who are being offered up to avoid corporate liability.

This is not to urge that there be more criminal convictions of corporations or international law judgments against the government. If the corporate or government conduct warrants it, there are provisions for both civil and criminal sanctions. This is to say that, if AI has anything to do with a wrongful death, liability should not be unfairly shifted to the individual operators. The after-the-fact shift to individual responsibility represented by the Tesla SEC filing and the DoD Directive, simply acknowledge the reality that AI is too flawed for a corporation or the government to proudly take responsibility for the AI choices. However, it is disingenuous to use that acknowledgment to unfairly shift responsibility to individuals for wrongful deaths where AI had a significant role in causation.

In actual practice, individual operators should not be allowed to say, "AI made me do it." Conversely, corporations or the government should not be allowed to say, "There was an individual operator so it does not matter that AI steered the trolly down the wrong track." This should be a fact based inquiry as to who had the agency—one, the other or both—to make the choices, if any, that were the proximate cause of death. If the AI incorrectly senses a pedestrian on the right shoulder and swerves into a full school bus on the other side of the road, the corporate defendant should not be allowed to say that there was an operator who takes all responsibility. Simliarly, the government that deploys a LAW that senses a threat from a nursery school and launches an attack because its AI computed that the nursery school presented a threat should not be able to point to the otherwise fairly passive operator to shift responsibility.

*Robert Sanger is a Certified Criminal Law Specialist (Ca. State Bar Bd. of Legal Specialization) and has been practicing as a litigation partner, now principal shareholder at Sanger Dunkle Law, P.C., in Santa Barbara for 50 years. Mr. Sanger is a Fellow of the American Academy of Forensic Sciences (AAFS). He is an Adjunct Professor of Law and Forensic Science at the Santa Barbara College of Law. The opinions expressed here are those of the author and do not necessarily reflect those of the organizations with which he is associated. ©Robert M. Sanger.*

## Endnotes

[1] Philippa Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, 5 Oxford Review 5-15 (1967) (hereinafter, "Foot, *Double Effect*").

[2] Professor Sandel's class session relating to the "trolley problem" is preserved on YouTube at: https://www.bing.com/videos/riverview/relatedvideo?q=michael+sandel+trolly+problem&mid=547770D09B19104855FA547770D09B19104855FA.

[3] See the discussion by Elizabeth Yuko in "How *The Good Place* Goes Beyond 'The Trolley Problem," The Atlantic Monthly (October 2017). Excerpts of the show are available on-line and the entire episode is available through NBC.

[4] For a philosophical and lighthearted account, see, e.g., Thomas Cathcart, The Trolley Problem or Would You Throw the Fat Guy Off the Bridge? (Workman Publishing, 2013).

[5] Abhirup Roy, Dan Levine and Hyunjoo Jin, *Tesla wins bellwether trial over Autopilot car crash*, Reuters (April 22, 2023).

[6] Sebastian Blanco, *Report: Tesla Autopilot Involved in 736 Crashes since 2019*, Car and Driver (June 13, 2023).

[7] David Shepardson, *US opens investigation into fatal Tesla crash in Virginia*, Reuters (August 10, 2023); Sebastian Blanco, *NHTSA Investigating Tesla Autopilot–Related Crashes with Emergency Vehicles*, Car and Driver (August 16, 2021).

[8] Robert E. Trager and Laura M. Luca, *Killer Robots Are Here—and We Need to Regulate Them*, Foreign Policy Magazine (May 11, 2022).

[9] Spencer Ackerman, *41 men targeted but 1,147 people killed: US drone strikes—the facts on the ground*, The Guardian (November 14, 2014). The problems persist today but the use of LAW's in the current conflicts are too raw to use as examples for the purposes of this article. Nevertheless, all participants in current conflicts should examine their consciences regarding the deployment of LAW's.

[10] An interesting student Note was just published while this column was in progress which makes an effort to discuss criminal liability for AI related harm using the "trolley problem" as vehicle for analysis: Jake Feiler, *The Artificially Intelligent Trolley Problem: Understanding Our Criminal Law Gaps in a Robot Driven World*, 14 Hastings Sci. & Tech. L.J. 1 (2023). The conclusion herein differs from the Note but the scholarship regarding the background of the problem is impressive.

[11] This quote and other references to Foot's original article are to Foot, *Double Effect, supra*. Note that this and other thought experiments are constructed by philosophers to isolate variables, thus uncertainty (which is a fact of reality) is stipulated away. See, Dennis V. Lindley, Understanding Uncertainty, (Wiley, 2014).

[12] The use of "fat man" in the original text would be regarded today as a form of body shaming. Substitutions have been suggested,

including the “large man,” “the heavy man,” the “man with a heavy backpack,” and “the man.”

[13] Intellectual history abandoned discussion of the cave explorers and became fixated on the brief mention of the trolley problem.

[14] Foot, *Double Effect*, supra.

[15] Inspired by events during World War II, such as the execution of all adult males, extermination of women and children in the Village of Lidice in reprisal for not surrendering those who killed a Nazi official. The questions raised by sacrificing a village for a few tested the Kantian categorical imperatives not to kill and not to use a person as a means to an end where the consequences were so disparate in terms of the amount of human suffering. The Kantian dilemma extends to the categorical duty not to lie, yet another problem in an effort to prescribe deontic mechanisms for behavior in actual situations. *See*, Michael Cholbi, *The Murderer at the Door: What Kant Should Have Said*, 79 Philosophy and Phenomenological Research 17 (2009).

[16] *R. v Dudley and Stephens*, 14 QBD 273 (1884).

[17] Michael Sandel, Justice: What’s the Right Thing to Do?, (Farrar, Straus and Giroux, 2009) 21-24.

[18] This analysis applies to the concept of self-driving vehicles in all present and anticipated iterations, including but not exclusively those produced by Tesla, Inc. which has gotten most of the press regarding liability. However, Waymo, Honda, Mercedes-Benz, and Nuro are also in the game with Ford and other manufacturers are making their way in.

[19] SAE J3016™ “Recommended Practice: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles,” (last revised May 3, 2021).

[20] DoDD 3000.09 (January 25, 2023) 1.

[21] See a thoughtful discussion in Chapter 11 of Katherine B. Forrest, When Machines Can Be Judge, Jury and Executioner: Justice in the Age of Artificial Intelligence (World Scientific, 2021).

[22] Tesla, Inc., United States Securities and Exchange Commission filing (Form 10-K) for the fiscal year ending December 31, 2022. Available electronically through the EDGAR data archives.

[23] DoDD 3000.09, (January 25, 2023) 6.

[24] Sydney J. Freedberg, Jr., *Artificial Intelligence, Lawyers and Laws of War*, Breaking Defense Newsletter, (April 23, 2021).

[25] Gregory C. Allen, *DOD Is Updating Its Decade-Old Autonomous Weapons Policy, but Confusion Remains Widespread*, Center for Strategic and International Studies (June 6, 2022).

[26] Penal Code sections 192(c)(2), 192(c)(1), 191.5(b), 191.5(a), respectively.

[27] *People v. Watson*, 30 Cal.3d 290 (1981).

[28] There are minimum guaranteed rights to all individuals under the 1949 *Geneva Convention* and there are special rights for specific protected persons, both combatant and noncombatant, in the context of international armed conflict. For instance, all civilians are protected from the effects of hostilities, including not being targeted to be attacked or killed under the *Fourth Geneva Convention and by Additional Protocol*, (Part IV, Section 1, Articles 48–51). Violations of the Geneva Convention Protocols, can subject an actor to criminal prosecution in the International Criminal Court.