Engaging Science, Technology, and Society 5 (2019), 40-60 DOI:10.17351/ests2019.260

Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction

MADELEINE CLARE ELISH[1]
DATA & SOCIETY RESEARCH INSTITUTE

Abstract
As debates about the policy and ethical implications of AI systems grow, it will be increasingly important to accurately locate who is responsible when agency is distributed in a system and control over an action is mediated through time and space. Analyzing several high-profile accidents involving complex and automated socio-technical systems and the media coverage that surrounded them, I introduce the concept of a *moral crumple zone* to describe how responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system. Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a highly complex and automated system may become simply a component—accidentally or intentionally—that bears the brunt of the moral and legal responsibilities when the overall system malfunctions. While the crumple zone in a car is meant to protect the human driver, the moral crumple zone protects the integrity of the technological system, at the expense of the nearest human operator. The concept is both a challenge to and an opportunity for the design and regulation of human-robot systems. At stake in articulating moral crumple zones is not only the misattribution of responsibility but also the ways in which new forms of consumer and worker harm may develop in new complex, automated, or purported autonomous technologies.

Keywords
autonomous vehicles; responsibility; machine learning; human factors; accidents; social perceptions of technology; self-driving cars, robot; human-in-the-loop; human-robot interaction

Introduction
On March 18, 2018, a self-driving Uber car struck and killed a pedestrian crossing her bike in the middle of an Arizona roadway. At the steering wheel of the putative "autonomous vehicle," a safety driver sat. Her job was to monitor the car's systems and take over in the event of an emergency. The safety driver now may face criminal charges of vehicular manslaughter (Somerville and Shepardon 2018). A devastating accident has forced the question that had been

[1] Madeleine Clare Elish, Email: mcelish@datasociety.net

only a hypothetical question circulating among lawyers and pundits: if a driverless car kills someone, who or what is to blame?

Questions having to do with the responsibility of various agents in complex, computational systems are not new. Such issues have been looked at from the diverse perspectives of law and policy (Calo, Froomkin and Kerr 2016) human factors engineering (Cummings 2006; Sheridan and Parasuraman 2005), systems design (Friedman 1997; Leveson 2011), ethics (Lin, Abney, and Bekey 2014; Coeckelbergh 2011; Bryson, Diamantis and Grant 2017) and the sociology of risk and innovation (Perrow 1984; Vaughn 1996). This article aims to be in conversation with these literatures and to contribute to ongoing public debates about AI and autonomous systems by providing a concept with which to think about the construction and attribution of responsibility in complex "intelligent" systems. Specifically, I articulate the concept of a *moral crumple zone* to describe how responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous system.[2] Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a highly complex and automated system may become simply a component—accidentally or intentionally—that bears the brunt of the moral and legal responsibilities when the overall system malfunctions. While the crumple zone in a car is meant to protect the human driver, the moral crumple zone protects the integrity of the technological system, at the expense of the nearest human operator. What is unique about the concept of a moral crumple zone is that it highlights how structural features of a system and the media's portrayal of accidents may inadvertently take advantage of human operators (and their tendency to become "liability sponges") to fill the gaps in accountability that may arise in the context of new and complex systems.

Building on previous work analyzing the history of aviation autopilot litigation in the 20th century (Elish and Hwang 2015),[3] this article calls attention to how incongruities between control and responsibility arise in complex systems, and how, in turn, such mismatches shape public and expert perceptions of responsibility. First, I present two well-known accidents involving complex socio-technical systems in which I argue moral crumple zones emerge, the partial nuclear meltdown at Three Mile Island and the crash of Air France Flight 447. My aim is not to provide new insight into the causes or circumstances surrounding the accidents. Rather, my aim is to call attention to a dynamic within complex systems that I argue will be increasingly relevant in the

[2] In this paper, I use the terms autonomous, automation, machine and robot as related technologies on a spectrum of computational technologies that perform tasks previously done by humans. A framework for categorizing types of automation proposed by Parasuraman, Sheridan and Wickens (2000) is useful for specifically analyzing the types of perceptions and actions at stake in autonomous systems. They define automation specifically in the context of human-machine comparison and as "a device or system that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator." This broad definition positions automation, and autonomy by extension, as varying in degree not as an all or nothing state of affairs. They propose ten basic levels of automation, ranging from the lowest level of automation involving a computer that offers no assistance to a human to the highest level of automation in which the computer makes all the decisions without any input at all from the human.

[3] Our conclusions are supported by similar work, most notably, Mindell 2015.

context of autonomous, robotic, and AI systems. The circumstances surrounding and the immediate responses to and media coverage of these accidents demonstrate how accountability appears to be deflected off of the automated parts of the system (and the humans whose control is mediated through this automation) and focused on the immediate human operators, who possess only limited knowledge, capacity, or control.

Following the discussion of these cases, I apply the concept of the "moral crumple zone," in which human operators take on the blame for errors or accidents not entirely in their control, to accidents involving self-driving cars. In this context, I call attention to the ways in which users and operators of such systems may be held responsible for failures in ways that obscure other human actors who may possess equal if not greater control over the behavior of a purportedly "autonomous" system. At stake in articulating "moral crumple zones" is not only the misattribution of responsibility but also the ways in which new forms of consumer and worker harm may develop in new automated technologies. The article concludes by discussing the implications of these dynamics for future policy and regulatory decisions.

The Accident at Three Mile Island

The partial nuclear meltdown at Three Mile Island is one of the most well-known industrial accidents to have occurred in the United States. While no deaths were attributed directly to the partial meltdown, the accident profoundly shaped the course of the civilian nuclear energy industry in the United States (Behr 2009; Walker 2009). The accident has also become a paradigmatic example of complex system failure, classically theorized by Perrow (1984) as a "normal accident," given the unavoidable possibility of the failure occurring given the complex socio-technical characteristics of the system. Before proceeding, I present an overview of the plant and its operations. I then provide an abbreviated description of the events leading up to the partial meltdown, and analyze responses to the accident, arguing that the operators became the moral crumple zone of the failed system.

The Three Mile Island Nuclear Generating Station is a nuclear power plant on the Susquehanna River ten miles southeast of Harrisburg, Pennsylvania's capital. It was the eighth and largest nuclear power plant to be built in the United States. It consists of two units and is still in operation. The first unit came online in the fall of 1974, and the second unit began commercial operation in December of 1978. Three months later, on March 28, 1979, the second unit sustained a partial core meltdown, the first nuclear disaster in the United States.

On a schematic level, a nuclear reactor, like the one at Three Mile Island (TMI), uses heat from nuclear fission to generate steam, powering a turbine, which generates electrical energy. In use at TMI is a Babock & Wilcox reactor, consisting of a forty by fifteen feet steel container with eight and half to twelve-inch-thick walls, inside of which is a nuclear core. Inside this core, uranium nuclei fission, controlled chain reactions that split atoms apart, occurs, releasing thermal energy that is then used to convert water into steam to power a turbine. Two sets of pipes are involved in the conversion of heat to steam. One set of pipes, the primary cooling water, is heated by circulating through the core and absorbing its heat. This primary water then travels through

steam generator tanks, filled with the secondary cooling water. The water heated by the reactor, the primary cooling water, does not come in direct contact with the water in the steam generator tanks, the secondary cooling water. The primary cooling water, like radiator coils, heats the secondary cooling water in the steam generator tanks by circulating through thousands of small tubes. The circulation of water in both sets of pipes is of critical importance. If the primary cooling water cannot absorb the heat from the core, the core will become too hot and will melt, releasing radioactive waste and radiation, as well as melting everything with which it comes in contact. Every aspect of the reactor system is precisely calculated and calibrated to maximize efficient heat transfer and to prevent the core from overheating. All safety systems exist in at least duplicate. Theoretically, when the system was originally designed, every risk was calculated, planned for, and addressed by the automated system.

All the pipes through which water circulates must be constantly maintained and cleaned to prevent buildup of foreign matter that could lead to malfunction. Various filters within the feedwater pipe system itself also perform sieving functions. In the early morning of March 28, one of these filters became clogged.[4] It would later come to light that these filters had consistently caused problems that the plant management had ignored.

At 4 am, in the middle of the 11 pm-7am shift, two maintenance workers were in the basement trying to fix a clogged pipe in a subsection of the system involved in purifying the secondary cooling water. Unintentionally, the workers choked off the flow of the entire feedwater system, preventing the secondary cooling water from circulating. This failure triggered a full shutdown of the reactor and turbine. Within the automated system, such a shutdown had been planned for adequately and further emergency automatic controls kicked in. Within seconds of the shutdown, auxiliary feedwater systems were activated that would cool the core. However, a relief valve designed to release pressure in the core had been triggered. The valve opened as designed, but the mechanism jammed, and the valve never closed, as it should have. Consequently, the cooling water intended to circulate drained out of the tank rapidly. Additionally, pipes that should have transported water to the tank had been rendered useless; two days earlier, a routine testing procedure of the valves in question had accidentally been left closed. The incorrect position of the valve was not linked to any indicators in the control room, and the mistake went unnoticed. Within minutes, the purported foolproof safety systems of the plant had failed and resulted in a common-mode failure, a term that denotes the failure of safety systems and a class of event with such remote probability that planning was unnecessary.

Unfortunately, further actions in the control room contributed to the failure of the safety systems. The operators, in the midst of multiple visual and audio error messages, misinterpreted the situation and relied on system readings linked to the open valve, assuming that this was an effect, not a cause, of the problem. Thinking there was too much water flowing, they shut off the remaining auxiliary pumps that had automatically been engaged, manually overriding the automatic safety system, another common-mode failure.

[4] This narrative of events is based on existing authoritative historical accounts by Walker (2004), Ellis (1979) and Ford (1981).

For more than sixteen hours, the reactor was not adequately cooled, and later reports showed that over a third of the uranium core melted. Much longer, and the meltdown could have been catastrophic. In the days and weeks following the accident, the extent of the damage and the potential of radioactive contamination were hidden from the public by plant management and the Nuclear Regulatory Commission (NRC). Numerous commissions and federal studies were tasked with evaluating what had gone wrong and providing recommendations for future action, including the President's Commission on the Accident at Three Mile Island.

Based on press releases from plant management, the governor's office, and the Nuclear Regulatory Commission (NRC), news coverage in the weeks and months following the accident focused on the role of operator error, generally referred to as human error. A *Los Angeles Times* front-page headline from April 11, 1979, less than two weeks after the meltdown, stated "Nuclear Accident Blamed Primarily on Human Error" (Toth 1979). Reporting on the official NRC report that was released two months later, one Associated Press headline read, "Human Error Cited in 3-Mile Accident" (Benjamin 1979). The first paragraph stated: "Operators of the Three Mile Island nuclear plant inadvertently turned what could have been a minor accident into a major one because they could not tell what was happening in the reactor." Only at the end of the article was it stated that the plant design made it especially hard to control and that "in general, control rooms… often are poorly designed and make it hard for operators to figure out what's going on during an abnormal event."

Without a doubt, actions taken by the plant operators led to the accident and exacerbated its severity. A maintenance worker two days prior had indeed left a valve closed after a testing procedure that should have been left open. It was steps taken by a maintenance worker to fix a clogged pipe that resulted in halting circulation in the feedwater pipes. And it was operators in the control room who overrode the final safety system, which would have engaged the remaining backup water system. But to focus on these actions as isolated events is like focusing on a detail in the foreground while missing the bigger picture.

For instance, the design of the control room played a central role in compounding human misinterpretations of mechanical failures. Designed as an automated system with limited human oversight, the physical conditions of the system were not adequately represented in the control interface (Rubinstein 1979; Sheridan 1992). For instance, there were no direct indicators of the level of cooling water in the steam generator tank. The automated system received this information (which had triggered the automatic shutdown), but the operators had to infer the amount of water from an auxiliary tank linked to pressure monitoring. During the accident, this tank remained full and provided incorrect information about the system to the operators. The operators made incorrect decisions because they had incorrect information. One of the central recommendations of the report was the requirement to focus on human factors engineering and the importance of human-computer interaction design (Kemeny et al. 1979).

Additionally, if the frame is expanded beyond those immediately present during the accident, errors followed directly from other systemic errors. It later came to light that the workers had been directed to test the valves and document the testing in a way that cut corners

and saved money and time for the plant managers. The maintenance of valves, specifically at TMI and also in nuclear plant facilities generally, was deemed to be overlooked and under-regulated by an official within the NRC (Omang 1979). Specifically, the clogged pipe in question had been generating issues for weeks prior, but plant management chose not to shut down the reactor. Compounding these circumstances, one must also take into consideration the organizational and power dynamics that may have prevented operators concerned with safety procedures, or unsure about what actions to take, in what has been described as a management climate that viewed regulations as empty bureaucratic hoops (Perrow 1984).

Focusing only on the agency of operators misses other sites of interaction and dimensions of control exercised by other actors involved in the system, from the designers of the interfaces to the plant managers who created the conditions within which the operators could act, to the regulators who maintained a blind-eye toward industry standards. It was this level of system complexity in which interactions were tightly coupled that Perrow (1984) points out as necessarily producing system failure in the form of a "normal accident" at TMI. His and later accounts (Walker 2004) emphasize the systemic causes that contributed to the accident, and underscore the incompleteness of understanding the accident as the result of human error.

Still, news coverage and later popular accounts of the accident positioned the operators as the moral crumple zone of the system. Even while the Kemeny report (1979) highlighted the reigning "mindset" at the plant and how "systemic" problems were the basis for the accident, the report emphasized the role of human failures in contrast to functioning technology, stating in the early pages of the report,

> We are convinced that if the only problems were equipment problems, this Presidential Commission would never have been created. The equipment was sufficiently good that, except for human failures, the major accident at Three Miles Island would have been a minor incident. (Kemeny et al. 1979: 8)

Although later modified, the narrative placing blame on the operators existed following the accident, and continued to exist even as expert reports complicated that narrative. In the opening minutes of a PBS *American Experience* (1999) documentary about the accident, Mike Gray, a prominent local journalist at the time, said, "If the operators had not intervened in that accident at Three Mile Island and shut off the pumps, the plant would have saved itself. They [the designers] had thought of absolutely everything except what would happen if the operators intervened anyway."

As Vaughan (1996) argues in her study of the NASA Challenger accident, media narratives and popular understandings of socio-technical accidents have significant power in shaping responses and interpretations of such events. In the dearth of official accident investigations and the news media cycle, public media coverage shapes how individuals make sense of an accident and who is responsible in ways that have implications for formal and informal processes of accountability. Even at the time, experts and commentators alike were

aware of the powerful role of the media in shaping knowledge about the accident at TMI (Christiansen 1979; Kemeny et al. 1979).

The partial nuclear meltdown at TMI is a useful example through which to see how a mismatch might emerge between the actual and the imagined or perceived control over a system that an operator may have. Even as their actions were constrained, their culpability was focused on and singled out. They became the moral crumple zone for the system's failure in the eyes of many.

The Crash of Air France Flight 447

Although the partial meltdown at Three Mile Island occurred several decades ago, the challenges of shared control between humans and machines that contributed to the accident remain essentially unsolved. Occurring nearly forty years later, the crash of Air France Flight 447 has also become a paradigmatic example of the vulnerabilities inherent in complex socio-technical systems. In this section, I present a brief overview of flight automation systems and describe the series of events leading up to the crash. In the following discussion, I extend my analysis from the crash itself to some of the underlying dynamics of automation that characterize modern flight automation design, articulating the ways in which these dynamics contribute to positioning pilots as moral crumple zones.

En route from Brazil to France in 2009, Air France Flight 447 crashed into the Atlantic Ocean killing all 228 people on board. One of the deadliest crashes in the last decades of civil aviation, the accident has been described as particularly tragic because the fatal error could have been easily fixed (Langewiesche 2014). Viewed in a different light, the circumstances of the accident provide another example of how human operators become moral crumple zones in complex system failures.

Airbuses are designed as a fly-by-wire system, referring to the complete automation of flight controls in the aircraft. Fly-by-wire systems are designed to be foolproof, primarily by prioritizing the computational capacities of on-board computers over human mechanic control. In a fly-by-wire aircraft, the pilot interfaces with a computer that in turn controls the aircraft through hydraulic or electric actuators. In previous generations of flight control, the movement of the pilot would be directly linked to the mechanical movements in the plane. Attempts to automate flight control are far from new and have been entwined with the development of manual flight since the Wright Brothers (Draper 1955). What is important to note is the relationship between the pilot and the aircraft and how automation mediates this in varying degrees and structures pilot action.

Airbuses operate within four flight control laws, including Normal Law and Alternate Law. When Normal Law is in effect, the decisions of the autopilot trump any action by the pilot. In theory and in practice this prevents pilots from making any moves, accidentally or incorrectly, that would rupture the flight envelope, the precise set of aerodynamic conditions that allow a more than 200-ton aircraft like the A330 to fly through the air. However, automated systems cannot be programmed to predict and plan for every single event that may ever occur at any

point in the future. This is as true for aviation autopilots as well as state-of-the-art machine learning techniques. So-called "edge-cases" exist, which combine factors and contexts that could not be anticipated. Most accidents are edge-cases. As both a practical response and liability shield, autopilots are certified to work as closed systems that do not work under every condition. I will return to the matters of boundaries and certifications in the discussion below.

Alternate Law, perhaps counterintuitively, refers to a mode in which primary control is in the hands of the pilot. It is in effect when parts of the computer system or autopilot are unable to work as designed, such as if a sensor reading is absent, and is characterized by the lack of various flight protections in place in Normal Law. For example, under Normal Law, the flight computer would override any actions that would result in an aerodynamic stall, which could result from an incorrect angle of attack, the degree at which the airplane wing meets the oncoming air. Under Alternate Law, the pilots have no such safety net.

After an on-time departure from Rio de Janeiro, the flight proceeded for one hour and forty minutes without incident.[5] In addition to the flight attendants, there were three pilots aboard who would rotate into the cockpit during the eleven-hour duration of the flight, with two in the cockpit at any given time. One of the pilots, the Pilot in Command, was the most senior pilot, and the others were both relatively young pilots who had spent the majority of their flight hours in Airbus aircraft in which pilots spend more time monitoring systems than actively controlling the aircraft.

About an hour and half into the flight, they encountered ice crystals that accumulated in the airplane's Pitot tubes, sensors that measure airspeed. Frozen, the Pitot tubes could not transmit airspeed indications, which the autopilot requires to function. Because the autopilot was receiving indications it sensed as false, the plane reverted to Alternate Law, and an alarm sounded altering the pilots to this shift. Soon another alarm sounded, indicating a deviation in planned altitude. One of the more junior pilots, likely panicked, pulled the stick back, perhaps instinctively, in an attempt to climb. A few seconds later, another warning sounded and a synthetic male voice pronounced, "STALL."

At this point, the pilots should have had enough knowledge and time to fix this relatively simple problem of recovery from an aerodynamic stall at high altitude. While counter-intuitive on the ground, it is a fundamental principle in flying that to recover from a stall, in which the aircraft speed is too slow and the angle of attack of the wings is too steep, the solution is to point the noise of the plane downward, decreasing the angle of attack and drag of the wings, increasing speed and thus recovering from the stall.

Instead of lowering the nose of the plane, the pilot pulled back on the control stick, raising the nose of the plane trying to climb. In the following minute, numerous alarms went off as the two junior pilots frantically tried to control the plane. Likely adding to their debilitating panic, alarm lights flashed and a menagerie of error warnings rang. The angle of attack at this point in the flight should have been around 3 degrees, with a stall occurring at 10 degrees. In

[5] This narrative is based on the journalistic account of the accident by Langewiesche (2014), as well the official BEA report (2012a).

their confused state, one of the pilots had brought the plane up as high as 23 degrees, and while the other tried to take over control, the design of the Airbus controls only allow one pilot to be in control at a time. The design also does not provide haptic feedback to indicate what the other pilot is doing, or even which pilot is in control if both are operating the controls. One pilot was pushing forward, the other pushing back. Neither was aware of the actions of the other. One minute and seventeen seconds had passed since the reversion to Alternate Law.

At this point, the plane was still above 30,000 feet and a recovery was theoretically easily within reach. But the chaos in the cockpit and breakdown in communication and coordination of the aircraft rendered all the pilots helpless, even though senior pilot had joined the other two in the cockpit by this point. The angle of attack had reached 41 degrees, so extreme that the computer did not announce a stall state because the reading was rendered invalid. Every time one of the pilots would lower the nose and reduce the angle of attack, the reading would fall back into the acceptable range, and a stall state would be announced. Any effectively correcting move he made perversely resulted in the synthesized male voice announcing "STALL," adding to the cacophony of other warnings. In the seconds before the crash, one of the pilots exclaimed, "We lost all control of the aeoroplane we don't understand anything we've tried everything" (BEA 2012b: 27). Four minutes and twenty seconds after the Pitot tubes froze, Flight 447 crashed into the Atlantic Ocean, killing everyone onboard instantly.

After the black boxes of the Airbus A330 were found in 2011, an accident investigation was completed by France's Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile (BEA), an equivalent body to the American Federal Aviation Administration (FAA). The report (BEA 2012a) concluded that the frozen Pitot tubes had set off the chain of events that caused the accident, although it was the series of responses by the crew that ultimately resulted in the crash. The report described how a combination of factors, including system feedback mechanisms, as well as insufficient crew responses, communication, and training were causes of the events leading up to the crash. However, American news outlets headlined the role of the pilots, focusing on the official French report's discussion of the pilots' inability to comprehend the situation and act in response. Many of the details described above were subsumed under a narrative in which the pilots lost "cognitive control," (BEA 2012a: 199) in the words of the BEA report, and caused the crash. A typical news report, here from CNN, explained,

> When ice crystals blocked the plane's Pitot tubes… the autopilot disconnected and the pilots did not know how to react to what was happening. In the first minute after the autopilot disconnection, the failure of the attempt to understand the situation and the disruption of crew cooperation had a multiplying effect, inducing total loss of cognitive control of the situation. (CNN 2012)

Buried in the second half of the story, it is explained that there were other factors involved in the crash, including the fact that Airbus had recognized an issue with Pitot tube failures due to icing in the A330 model, and were beginning to replace the parts. The pitot tubes on this particular Airbus A330 had not yet been replaced.

It is interesting to contrast this narrative with the marketing and reporting around an early model of the A330, the Airbus A320, the first fly-by-wire commercial jet. Quoting an aviation expert, the article reporting on the model's debut states,

> ...most significant is that computers controlling the fly-by-wire system can be programmed to ensure that the plane flies safely at all times, even though the pilot may make an error. … It will be smart enough to protect the airplane and the people aboard it from any dumb moves by the pilot. (Oslund 1986)

The explicit point in this article, as well as similar media from the time, is that the autopilot and associated automation are smart enough to outsmart and save the human every time, the same narrative we saw in nuclear power plant design. The idea that the automation and its software could fail was never a possibility.

If the software is presented as being more capable of control, and the amount of time on any given flight that is controlled by the autopilot software far exceeds the amount of time directly controlled by the pilot, who is responsible for the control of the aircraft? The FAA has specifically addressed this in a federal regulation, which has been the same for decades: "The pilot in command of an aircraft is directly responsible for, and is the final authority as to, the operation of that aircraft" (14 CFR 91.3). Courts have consistently upheld this authority of the pilot as the ultimate designation of liability (Cooling and Herbers 1983). While control has been effectively distributed, responsibility has not scaled accordingly.[6]

Historians of technology have demonstrated in a variety of contexts and in a variety of time periods that it is a social tendency to overestimate the capacity of machines and underestimate the abilities of humans (Mindell 2015). In addition, this has been a sustained frame for analyzing accidents in a Western historical context (Barnaby 1968), and pilot error has been a consistent catchall for explaining commercial and private aircraft accidents (Leveen 1982; Elish and Hwang 2015).[7] It is of course reasonable to hold humans accountable because non-human entities cannot be held as accountable to society in ways that contribute to justice and the greater public good. But when "human error" is invoked, it generally refers to operator error, not the error of human designers or systems architects.

These explanatory tendencies are insufficient for accounting for the complex and distributed agency within aviation human-computer systems. Regulators, in addition to the engineers and managers of aviation systems, have created contradictory dynamics in which automation is seen as safer and superior in most instances, unless something goes wrong, at

[6] This argument is not intended to be against automated systems in and of themselves. The safety record in aviation over the past decades demonstrates that highly automated systems have resulted in significantly safer air travel overall.

[7] A leader in aviation accident investigations, Jerome Lederer (1974) took a position against the prevailing one held by the NTSB, arguing that classifications of pilot error do not explain *why* an accident occurred. Instead, he insisted that it was necessary to use "categories that would acknowledge the interactions between humans and machines, such as a pilot error induced by design of aircraft, error as a result of ignorance, error due to deliberate acts not in accordance with good practice, error caused by environment, and error caused by psychological or social reasons."

which point humans are regarded as safer and superior. Unfortunately, creating this kind of role for humans, who must jump into an emergency situation at the last minute, is something humans do not do well (Roscoe 1992, Weiner 1989), as was the case in the Air France crash.

Indeed, decades of engineering human factors research helps explain how the Air France pilots in many ways were primed to "lose cognitive control of the situation." While automation is generally assumed to relieve humans of menial tasks, freeing them to think about more important decisions, this has proven not to be the case (Bainbridge 1983; Parasuraman and Riley 1997). More free time does not necessarily lead to high-level judgments. In fact, pilot awareness generally decreases with increased automation (Casner and Schooler 2014). Human factors research has demonstrated that skills atrophy when automation takes over (Sarter et al. 1997). While the senior pilot had experience flying a range of aircraft, the other two pilots had much less experience and had only flown for a significant amount of time in fly-by-wire Airbuses. Deskilling has been suggested to be a primary component of the pilots' inability to implement the stall corrective procedure (Langewiesche 2014).

Moreover, the framework of autopilot certification bounds the automatic system in a way that limits accountability to only mechanical failure. The autopilot is functioning correctly, according to certification standards, as long as the human pilot is provided the specified amount of time to take control in the event of an accident (FAA 2011 AC 23-17C). Human factors research has proven this "handoff" scenario detracts from, rather than enhances, human performance. The autopilot system is certified as a piece of software, but in practice works as an interactional human-software-hardware system. If, as in Flight 447, the primary causes of the accident are found in the interactions *between* automation and human, there are no certifications that cover this. Because the autopilot did not malfunction in a way recognized through its certification process, the only possible malfunction, systemically, is the human pilot, becoming the moral crumple zone.

Discussion

In an article titled "Accountability in a Computerized Society," Helen Nissenbaum (1996) outlined four main barriers to the establishment of accountability, or what she termed answerability, in the development and use of computational technologies. Each of these barriers (the problem of many hands, bugs, blaming the computer, and software ownership without liability) implicates a set of development practices as well as a set of social attitudes toward accountability. She argues that computational technologies create gaps in accountability in ways that are systemic and necessary to address. The concept of a moral crumple zone is not a way to explain why these gaps occur. Rather, the concept highlights how human operators may "absorb" responsibility when these gaps arise in socio-technical systems in ways that do not reflect the distributed control and interactional aspects that compose the socio-technical system.

However, moral crumple zones are not a property of every complex and automated system. There may be some instances where organizational structures or egregious product defects prevent the misattribution of blame. For instance, in the mid-1980s, numerous lawsuits

were brought against the manufacturer of the Therac-25, a computerized radiation therapy machine. There were six known accidents involving massive overdoses of radiation delivered by the machine. The accidents occurred when the technician operating the machine rapidly entered an incorrect series of commands that triggered the machine to physically release a low-dose of radiation but to represent an error state to the technician, indicating that the dose of radiation had not been delivered. The error, which resulted in the technician's delivering multiple doses of radiation, was proven to be a software error, and not the result of technician error. In the press, a *New York Times* headline attributed the error to "Computer Mistake," and the opening paragraph explained, "A computer malfunction apparently caused excessive radiation doses for two cancer patients at a treatment center, causing the death of one man…" (AP 1986).

As is the case with all complex systems, the causes of accidents are multiple and pointing to one error is usually a vast overstatement of the problem (Leveson and Clark 1993). Indeed, Nissenbaum (1996) uses the Therac-25 accidents as an example of the "the problem of many hands," and describes how the plethora of actors, from multiple computer programmers to corporate executives involved in the development of Therac-25, obscures the responsibility of key individuals.

This counterexample brings into relief the characteristics of systems or accidents in which we might expect to see moral crumple zones emerge. First, moral crumple zones are likely to take shape in the immediate aftermath of a highly publicized event or accident. It is notable that the Therac-25 accidents became public nearly two years after the first overdose of radiation was given. The malfunction and harm was not immediately clear, but rather came to light slowly. In other words, when the accidents were reported, months of inquiries into the causes of the error had taken place, and fault had already been assigned to the manufacturer. Moreover, media coverage plays a significant role in shaping social perceptions of responsibility and accountability and in creating the context in which moral crumple zones emerge and take hold.

Second, we can see a difference in the position in which the operator of the system was placed. In the case of Therac-25, the operator had no way of knowing that the system had malfunctioned, except for reports from patients that felt pain. In the case of the TMI meltdown, the operators knew the system was malfunctioning, but they did not have sufficient information or authority to take corrective actions. In the case of the Air France crash, the extent to which the system was malfunctioning was only variously visible, and while the pilots had sufficient information to react, in many ways they were systemically disempowered to act appropriately on that information

Moreover, the Therac-25 failed to perform as designed, and this failure to carry out the intended action resulted in a radiation overdose. In the case of TMI, the system performed as designed—but ultimately it was a failure of how the system, itself, was designed and maintained that created the conditions for a partial meltdown to take place. In the case of the Air France crash, once again, the system performed as designed in the face of a mechanical failure and still the accident occurred. This dynamic underscores how a moral crumple zone is more than just the articulation of a scapegoat. The term is meant to call attention to the ways in which automated and autonomous systems deflect responsibility in unique and structural ways,

protecting the integrity of the technological system at the expense of the nearest human operator. The technology is maintained as faultless, while the human operator becomes the faulty feature of the system.

This article has focused not on legal liability but rather on cultural perceptions of blame and responsibility, particularly in an American context. Unfortunately, it is beyond the scope of this article to explore specific legal theories of liability. Still, I hope to have demonstrated that the cultural perceptions of fault in automated and robotic systems may permeate formal frameworks of accountability through both media accounts and official accident reports. Especially in the context of emerging technologies, social norms and expectations play a significant role in the legal integration of a technology into existing frameworks. For instance, perceptions of new technologies become condensed in the metaphors used to describe technology and its effects. These metaphors influence the outcome of legal interpretations of new technology (Froomkin 1995; Calo 2016).

Framing cultural perceptions of accountability in the context of moral crumple zones can provide a means to think about how risk and responsibility is—or should be—distributed in socio-technical systems. With regard to autonomous and robotic technologies, the regulations, laws, and norms are still in formation, and may be particularly susceptive to uncertainties or even evasions of responsibility (Graham 2012). Additionally, societal expectations around these technologies may prevent people from leveraging their legal rights, if they believe they are at fault. The concept of the moral crumple zone is useful in thinking through the instances in which unfairness or harm might arise but that are not yet formally addressed or even recognized.

Robots on the Road

In this section I bring forward other scenarios in which we might see moral crumple zones emerge, including the recent case of the 2018 self-driving Uber car accident that was described at the beginning of this article. Self-driving cars are likely to be one of the first intelligent and semi-autonomous technologies to be widely adopted. We have yet to see all the ways in which liability will, or will not, be distributed. Will self-driving cars create moral crumple zones? Probably.

When news of the self-driving Uber car accident first surfaced, media coverage focused on the consequence of the accident: the death of a pedestrian, with headlines like: "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam" (Wakabayashi 2018) and "Self-driving Uber car hits, kills pedestrian in Tempe, Arizona crash" (ABC15 2018). In the days following the crash, video footage from the car was released. One set of footage, with the camera positioned facing the safety-driver, showed the final moments before the crash, with the safety-driver sitting still and glancing down into her lap until, presumably, seeing the pedestrian, she gasps in horror. The second set of footage, with a camera positioned facing the road, captured the approach toward the pedestrian and moment of impact. These two gruesome clips were looped and repeated on websites and American TV. Headlines shifted, reasonably, to reporting on potential causes of the crash. And relatively quickly, media coverage began to focus on the safety-driver as the cause of the accident. As the Tempe, Arizona Chief of Police told one reporter

in an interview: "I suspect preliminarily it appears that the Uber would likely not be at fault in this accident… I won't rule out the potential to file charges against the [backup driver] in the Uber vehicle" (Said 2018).

An NTSB preliminary report (2018), released two months after the crash, detailed the software failures that led to the accident. Two different types of software failures were implicated. The system used to detect and classify objects around the car misrecognized the pedestrian as an object. In addition, software that might have enabled automatic braking had been disabled: the self-driving car was a modified Volvo XC90 SUV, equipped with many driver-assistance features, but running Uber's own self-driving software. When the Uber car's software is in autonomous mode, the safety features of the Volvo are disabled. Had this not been the case, it is expected that the Volvo would have engaged the brakes and stopped before hitting the pedestrian (Lee 2018).

The report and subsequent media coverage also cited the safety driver's failure to brake in time, and while the safety driver may have been looking down at the touch-screen used to monitor the self-driving car systems, concerns were also raised as to whether the safety-driver was looking at her cell phone or streaming media (O'Brien 2018).

The circumstances leading to the crash involved a complex set of factors, including software flaws and the role of the safety driver. Given the known existence of the "hand-off problem," described in the aviation context above, it is reasonable to question the appropriateness of the role and expectations of the safety driver in and of itself. Nevertheless, it is the safety driver who may now be facing criminal charges for vehicular manslaughter (Somerville and Shepardson 2018).

In addition to Uber, Tesla also continues to develop self-driving cars in which the human backup driver is at once superfluous and essential. Consider, for instance, a potential feature of Tesla's self-driving car. In 2015, Tesla proposed that if a car were going to switch lanes in autonomous mode, a human would have to "sign off" on the lane change by clicking on a turn signal indicator presented to the operator (Ramsey 2015). Elon Musk, referring to a new release of Tesla Autosteer software that year, warned:

> It's almost to the point where you can take your hands off [. . .] but we're very clearly saying this is not a case of abdicating responsibility…. The hardware and software are not yet at the point where a driver can abdicate responsibility…. [The system] requires drivers to remain engaged and aware when Autosteer is enabled. Drivers must keep their hands on the steering wheel. (Sorokanich 2015)

While elsewhere the autonomy of the Tesla Autosteer is emphasized, here we see how the human retains all responsibility. It is clear to see the parallels to the paradigm of "human in the loop" supervised automation that has developed in aviation.[8]

[8] See Stilgoe (2017) for a related analysis of the Tesla 2016 crash in which a driver was killed while his car was in Autopilot mode. Stilgoe observes Tesla and the NTSB as pushing responsibility onto drivers in all

In contrast, Google designers seem by and large aware of the pitfalls that surround supervised automation. Google's self-driving car program has switched focus after making the decision that it could not reasonably solve the "handoff problem," that is, having the car handle all the driving except the most unexpected or difficult situations (Markoff 2016).

Nonetheless, intelligent and autonomous systems in every form have the possibility to generate moral crumple zones because they distribute control, often in obfuscated ways, among multiple actors across space and time. This is especially the case, in the current context of self-driving cars in which frameworks for responsibility and liability are underdetermined (Graham 2012), even as these technologies are being tested on streets and marketed to consumers as if they have already been determined as safe and successful (Stilgoe 2017).

Another example might be seen in the current discourses around Google's driverless car accidents. Between 2015 and 2017, Google made public the accident record of its self-driving car tests (Kovach 2017). The periodic announcements and subsequent press coverage declared that all except one of the accidents had been caused by the Google car; all were the fault in some way of human drivers.

Still, there was a surprising pattern of rear-end accidents (Davies 2016). Perhaps these kinds of accidents are the most common on the stop-and-go streets of Palo Alto.[9] It is also possible that the Google car effectively caused some of the accidents in that it was driving in a way contrary to the expectations of the drivers around it. Driving is as much about reacting to other drivers, being able to anticipate what they are likely to do, as it is about obeying stop signs and avoiding obstacles. Maybe the Google car is more cautious or slow than most drivers in the area, and so the human drivers anticipated the car's movement incorrectly. The accidents might have been caused by a fundamental miscommunication between a driverless car and a human-driven car. In this instance, responsibility is shifted to other drivers on the road, and these human drivers become the moral crumple zone, taking on responsibility for a failure where, in fact, control over the situation is shared.

Identifying the boundaries of actors within systems of shared control can be tricky. Where does the agency of the engineer end and the operator begin? How does one delineate the boundaries of a system that is necessarily socio-technical? In this differentiation, there are significant consequences for how each actor may be held accountable. Technology safety certifications are one way in which the boundaries of actors have been established.

As described above in the context of autopilots, certifications can be a means to track agency in distributed systems and investigate accountability. However, current paradigms of certifications do not take into account the interactional aspect of system components. How might certifications be reframed to reflect the growing body of knowledge within the human factors community about human-machine interaction? Moreover, issues of certification will most

---

control modes, concluding, "The identification and blaming of human deficits has been a common feature of self-driving car innovation" (41).

[9] Research has drawn attention to the limitations of assessing with certainty the safety of driverless cars on the road, especially when safety metrics are drawn in comparison to traditional vehicles (Schoettle and Sivak 2015; Kalra and Paddock 2016).

certainly come up in regards to deep learning technologies, and other emergent forms of artificial intelligence (Umson 2016). How do you certify what is an unbounded system? Such questions would be productive areas of future research.

Conclusion

This article has proposed the concept of a moral crumple zone as a provocation to rethink how, why, and with what implications responsibility will be assigned when automated, autonomous, or "intelligent" systems fail. The concept is a way to account for the incongruities between control and responsibility that may arise when control over an action or function has become distributed across multiple actors (human and nonhuman), and the implications of these incongruities for social perceptions and formal structures of responsibility in intelligent systems.

As debates about the policy and ethical implications of AI systems grow, it will be increasingly important to accurately locate who is responsible when agency is distributed in a system and control over an action is mediated through time and space. When humans and machines work together, traditional conceptions of control and responsibility will likely need to change in response. This is an especially pressing question given that recent reports on the future of work and automation emphasize that computers will not replace workers, but rather help workers do their jobs better (Chui et al. 2015; Davenport and Kirby 2016). A prevailing rhetoric of human-computer interaction design suggests that keeping a "human in the loop" assures that human judgment will always be able to supplement automation as needed. This rhetoric emphasizes fluid cooperation and shared control. In practice, the dynamics of shared control between human and computer system are more complicated, especially with respect to issues of formal mechanisms of accountability (Jones 2015).

This article has attempted to articulate a problem and characterize a set of frictions that emerge when automated systems disrupt traditional linkages between control and responsibility. The discussion has ultimately been two-fold. In the first part, I articulated the potential mismatches that can occur between control and responsibility in automated systems through a discussion of the nuclear meltdown at Three Mile Island and the crash of Air France Flight 447. These mismatches, I argued, create moral crumple zones, in which human operators take on the blame for errors or accidents not entirely in their control. In the final part of the article, I brought the idea of the moral crumple zone out of the context of industrial systems and asked what it might look like in the context of commercial technologies, namely, driverless cars. I also explored how traditional modes of technology certification may reify the potential to create moral crumple zones and suggested that a reexamination of certification paradigms may be a productive avenue of future research.

This article presents the concept of the "moral crumple zone" as both a challenge to and an opportunity for the design and regulation of human-robot systems. At stake in the concept of the moral crumple zone is not only how accountability may be distributed in any robotic or autonomous system, but also how the value and potential of humans may be allowed to develop in the context of human-machine teams.

Author Biography

Madeleine Clare Elish is a Research Lead and co-founder of the AI on the Ground Initiative at Data & Society Research Institute, an independent non-profit research institute focused on the social implications of data-centric technological development. She received her PhD in Anthropology from Columbia University, and an M.S. in Comparative Media Studies from MIT.

Acknowledgements

I am grateful to the many who provided feedback and encouragement at various stages of this article's development. My deepest thanks to Tim Hwang and Robin Sloan, danah boyd, and Andrew Selbst, as well as to Rebecca Crootof, Sue Glueck, Ian Kerr, Ryan Calo, Woody Hartzog, Michael Froomkin and the WeRobot conference community, where a previous version of this paper was presented in 2015. I also thank Katie Vann, Daniel Lee Kleinman and the three anonymous peer-reviewers whose comments and suggestions greatly strengthened this article. This research was supported in part by the John D. and Catherine T. MacArthur Foundation and the Ethics and Governance of AI Fund.

References

ABC 15 staff. 2018. "Self-driving Uber car hits, kills pedestrian in Tempe." *ABC 15 Arizona*, Mar 19. Accessed 1 July 2018. https://www.abc15.com/news/region-southeast-valley/tempe/tempe-police-investigating-self-driving-uber-car-involved-in-crash-overnight.

AP, 1986. "Fatal Radiation Dose in Therapy Attributed to Computer Mistake." *New York Times* 21 June, pg. 50. Accessed 1 July 2018. https://www.nytimes.com/1986/06/21/us/fatal-radiation-dose-in-therapy-attributed-to-computer-mistake.html.

Bainbridge, L. 1983. "Ironies of Automation." *Automatica* 19:775-779.

Barnaby, K.C. 1968. *Some Ship Disasters and Their Causes*. Cranbury N.J.: A.S. Barnes.

BEA, 2012a. *Final Report on the Accident on 1st June 2009*. Accessed 1 July 2018. https://www.bea.aero/en/investigation-reports/notified-events/detail/event/accident-de-lairbus-a330-203-immatricule-f-gzcp-et-exploite-par-air-france-survenu-le-01062009-da/.

BEA. 2012b. *Final Report on the Accident on 1st June 2009*: *Appendix 1:CVR Transcript (EN)*. Accessed 1 July 2018. https://www.bea.aero/uploads/tx_elyextendttnews/annexe.01.en_03.pdf.

Behr, P., 2009. "Three Mile Island Still Haunts U.S. Nuclear Power Industry," *New York Times*, March 27. Accessed 1 July 2018. http://www.nytimes.com/gwire/2009/03/27/27greenwire-three-mile-island-still-haunts-us-reactor-indu-10327.html.

Benjamin, S., 1979. "Human Error Cited in 3-mile Accident." *Boston Globe.* May 12, pg 5.

Bryson, J. J., M. E. Diamantis, T.D. Grant. 2017. "Of, for, and by the people: the legal lacuna of synthetic persons." *Artificial Intelligence Law* 25: 273-291.

Calo, R 2016. Robots in American Law. We Robot 2016 presentation. University of Washington School of Law Research Paper 2016-04. Accessed 17 Feb 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2737598.

Calo, R. A. Michael Froomkin, and Ian Kerr. eds. 2016. *Robot Law*. Northampton: MA: Elgar Publishing.

Casner, S.M. and J. Schooler, 2014. "Thoughts in Flight: Automation Use and Pilots' Task-Related and Task-Unrelated Thought." *Human Factors* 56(3,):433-422.

Chui, M. J. Manyika and M. Miremadi, 2015. "Four Fundamentals of Workplace Automation." *McKinsey Quarterly* November. Accessed 1 July 2018. http://www.mckinsey.com/business-functions/business-technology/our-insights/four-fundamentals-of-workplace-automation (accessed 1/2/2016);

CNN, 2012. "Final Air France crash report says pilots failed to react swiftly" *CNN.com* July 5. Accessed 1 July 2018. http://www.cnn.com/2012/07/05/world/europe/france-air-crash-report/.

Code of Federal Regulations, 14 CFR 91.3.

Coeckelbergh, M., 2011. "Moral Responsibility, Technology, and Experiences of the Tragic: From Kierkegaard to Offshore Engineering." *Science and Engineering Ethics* 18(1):35-48.

Cooling, J.E. and P. V. Herbers. 1983. "Considerations in Autopilot Litigation." *Journal of Air Law and Commerce* 48: 693-723.

Christiansen, D. 1979. "TMI and the press." *IEEE Spectrum* November: 92-95.

Cummings, M.L. 2006. "Automation and accountability in decision support system interface design." *Journal of Technology Studies* 32: 23-31.

Davenport, T. and Kirby, J. 2015. "Beyond Automation." *Harvard Business Review* 93 (6). Accessed 1 July 2018. https://hbr.org/2015/06/beyond-automation.

Davies, Alex. 2016. "Google's self-driving car caused its first crash." *Wired* Feb 29. Accessed 17 Feb 2019 https://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash/

Draper, C.S, 1955. "Flight Control" 43rd Wilbur Wright Memorial Lecture. *Journal of the Royal Aeronautical Society* 59 July: 451-478

Elish, M.C. and Tim Hwang. 2015. "Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation." May 18. Data & Society Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2720477

FAA. 2011. AC 23-17C - Systems and Equipment Guide for Certification of Part 23 Airplanes and Airships. Accessed 1 July 2018. https://www.faa.gov/regulations_policies/advisory_circulars/index.cfm/go/document.information/documentID/1019689 .

Ford, D.,1981. "A Reporter at Large: Three Mile Island." *The New Yorker* Apr 6: 49-120.

Friedman, Batya, ed. *Human Values and the Design of Computer Technology*. New York: University of Cambridge Press.

Froomkin, M., 1995. "The Metaphor is the Key: Cryptography, the Clipper Chip, and the Constitution." *U. PA. L. Rev.* 143(3): 709-862.

Graham, K. "Of Frightened Horses and Autonomous Vehicles: Tort Law and its Assimilation of Innovations," *Santa Clara L. Rev*. 52: 1241.

Jones, M. L. 2015. "The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practice Principles. *Vand. J. Ent. & Tech*. 18: 77-134.

Kovach, S.. 2017. "Google quietly stopped publishing monthly accident reports for its self driving cars." *Business Insider* Jan 18. Access 17 Feb 2019. https://www.businessinsider.com/waymo-ends-publishing-self-driving-car-accident-reports-website-2017-1

Kalra, N. and S.M. Paddock. 2016. Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability. RAND Corporation, Santa Monica, CA. Accessed 1 July 2018. www.rand.org/t/RR1478.

Kemeny, J.G. et al., 1979. "Report of the President's Commission on the Accident at Three Mile Island," U.S. Government Printing Office, 0-303-300, October.

Langewiesche, W., 2014. "The Human Factor." *Vanity Fair* September. Accessed 1 July 2018. http://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash.

Lederer, J. 1974. "Human Factors & Pilot Error," *Air Line Pilot* July:13-14.

Lee, T. 2018. "NTSB: Uber's sensors worked; its software utterly failed in fatal crash." *ArtsTechnica*, June 24. https://arstechnica.com/cars/2018/05/emergency-brakes-were-disabled-by-ubers-self-driving-software-ntsb-says/ (accessed 7/1/2018).

Leveen, S.A., 1982. "Cockpit Controversy: The Social Context of Automation in Modern Airlines." Ph.D. Dissertation, Department of Science and Technology Studies, Cornell University.

Leveson, N.G. and C.S. Turner, 1993. "An Investigation of the Therac-25 Accidents," *IEEE*, July:18-41.

Leveson, N. G.. 2011. *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge, MA: MIT Press.

Lin, P., K. Abney, and G. A. Bekey, eds. 2014. *Robot ethics: The ethical and social implications of robotics.* Cambridge, MA: MIT Press.

Markoff, J., 2016. "Google Car Exposes Regulatory Divide on Computers as Drivers." *New York Times*, Feb10. http://www.nytimes.com/2016/02/11/technology/nhtsa-blurs-the-line-between-human-and-computer-drivers.html?smid=tw-share&_r=0 (accessed 2/11/2016).

Mindell, D. A. 2015. *Our Robots, Ourselves*. New York: Viking.

Nissenbaum. H., 1996. "Accountability in a Computerized Society."*Science and Engineering Ethics* 2(1):25-42.

NTSB. 2018. HWY18MH010 – Preliminary Report. National Transportation Safety Board (NTSB), May 24. Accessed 1 July 2018. https://goo.gl/2C6ZCH,

O'Brien, S.A. 2018. "Uber operator in fatal self-driving vehicle crash was likely streaming 'The Voice." *CNN*, Jun 22. Accessed 1 July 2018. http://money.cnn.com/2018/06/22/technology/uber-self-driving-crash-police-report/index.html.

Omang, J. 1979. "'Nuggets': A Collection of Nuclear Glitches." *Washington Post*, Feb 10. Accessed 17 Feb 2019. https://www.washingtonpost.com/archive/politics/1979/02/10/nuggets-a-collection-of-nuclear-glitches/a48dbfae-d6d8-45ef-964c-a433a5f6bdf6/.

Oslund, J., 1986. "NWA Airbus 320s to be most advanced jets ever." *Minneapolis Star Tribune*. 9 Oct.

Parasuraman, R. and V. Riley, 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* June 39(2): 230-253.

Parasuraman, R., T.B. Sheridan, C.D. Wickens. 2000, "A Model for Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man and Cybernetics*, volume 30, issue 3 (May) pp. 286-297.

Perrow, C.1984. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books.

PBS, 1999. "Meltdown at Three Mile Island." *PBS American Experience*. Chana Gazit, dir. WGBH Boston.

Ramsey, M. 2015. "Who's Responsible When a Driverless Car Crashes? Tesla's Got an Idea." *Wall Street Journal*, May 13. Accessed 1 July 2018. http://www.wsj.com/articles/tesla-electric-cars-soon-to-sport-autopilot-functions-such-as-passing-other-vehicles-1431532720.

Roscoe, A.H., 1992. *Workload in the glass cockpit. Flight safety digest.* Alexandria, VA: Flight Safety Foundation

Rubinstein, E. 1979. "The accident that shouldn't have happened." *IEEE Spectrum* November: 33-57.

Said, C. 2018. "Exclusive: Temple Police chief says early probe shows no fault by Uber." *San Francisco Chronicle* March 26. Accessed 1 July 2018. https://www.sfchronicle.com/business/article/Exclusive-Tempe-police-chief-says-early-probe-12765481.php.

Sarter, N.B., D.D. Woods, and C.E. Billings, 1997. "Automation Surprises," In: G. Salvendy, ed. *Handbook of Human Factors & Ergonomics 2nd ed*, New York: Wiley.

Schoettle, B. and M. Sivak, 2015. *A Preliminary Analysis of Real-World Crashes Involving Self-Driving Vehicles*. October. University of Michigan: Transportation Research Institute. UMTRI-2015-34. Accessed 1 July 2018. http://www.umich.edu/~umtriswt/PDF/UMTRI-2015-34_Abstract_English.pdf.

Sheridan, T. 1992. *Telerobotics, automation and supervisory control.* Cambridge, MA: MIT Press.

Sheridan, T. B. and R. Parasuraman. "Human-Automation Interaction." *Review of Human Factors and Ergonomics* 1(1): 89-129.

Somerville, H. and D. Shepard. 2018. "Uber car's 'safety' driver streamed TV show before fatal crash: police." *Reuters*, June 22. Accessed 1 July 2018. https://www.reuters.com/article/us-uber-selfdriving-crash/uber-cars-safety-driver-streamed-tv-show-before-fatal-crash-police-idUSKBN1JI0LB.

Sorokanich, B. 2015. "Tesla Autopilot First Ride." *Road & Track*, Oct 14. Accessed 1 July 2018. http://www.roadandtrack.com/new-cars/car-technology/news/a27044/tesla-autopilot-first-ride-almost-as-good-as-a-new-york-driver/.

Stilgoe, J. 2017. "Machine learning, social learning and the governance of self-driving cars." *Social Studies of Science 48*(1), 25–56. https://doi.org/10.1177/0306312717741687

Toth, R. 1979. "Nuclear Accident Blamed Primarily on Human Error." *Los Angeles Times* Apr 11, pg. 1

Umson, C., 2016. Letter. NHTSA. Accessed 1 July 2018. http://isearch.nhtsa.gov/files/Google%20--%20compiled%20response%20to%2012%20Nov%20%2015%20interp%20request%20--%204%20Feb%2016%20final.htm

Vaughan, D. 1996. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago, IL: University of Chicago Press.

Wakabayashi, D. 2018. "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam." *The New York Times*, Mar 19. Accessed 1 July 2018. https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html.

Walker, J. S. 2004. *Three Mile Island: A Nuclear Crisis in Historical Perspective.* Berkeley, CA: University of California Press.

Walker, J. S. 2009. "Documenting Three Mile Island." *Bulletin of the Atomic Scientists.* March 18. https://thebulletin.org/2009/03/documenting-three-mile-island-2/ Accessed Feb 17 2019.

Weiner, E.L., 1989. *Human factors of advanced technology ("glass cockpit") transport aircraft* (NASA Contractor Report 177528). Moffett Field, CA: NASA Ames Research Center.