MORAL CRUMPLE ZONES:

CAUTIONARY TALES IN HUMAN-ROBOT INTERACTION

M.C. ELISH

COLUMBIA UNIVERSITY AND DATA & SOCIETY RESEARCH INSTITUTE

Abstract

A prevailing rhetoric in human-robot interaction is that automated systems will help humans do their jobs better. Robots will not replace humans, but rather work alongside and supplement human work. Even when most of a system will be automated, the concept of keeping a "human in the loop" assures that human judgment will always be able to trump automation. This rhetoric emphasizes fluid cooperation and shared control. In practice, the dynamics of shared control between human and robot are more complicated, especially with respect to issues of accountability.

As *control* has become distributed across multiple actors, our social and legal conceptions of *responsibility* remain generally about an individual. If there's an accident, we intuitively—and our laws, in practice—want some*one* to take the blame. The result of this ambiguity is that humans may emerge as "moral crumple zones." Just as the crumple zone in a car is designed to absorb the force of impact in a crash, the human in a robotic system may become simply a component—accidentally or intentionally—that is intended to bear the brunt of the moral and legal penalties when the overall system fails.

This paper employs the concept of "moral crumple zones" within human-machine systems as a lens through which to think about the limitations of current frameworks for accountability in human-machine or robot systems. The paper examines two historical cases of "moral crumple zones" in the fields of aviation and nuclear energy and articulates the dimensions of distributed control at stake while also mapping the degree to which this control of and responsibility for an action are proportionate. The argument suggests that an analysis of the dimensions of accountability in automated and robotic systems must contend with how and why accountability may be misapplied and how structural conditions enable this misunderstanding. How do non-human actors in a system effectively deflect accountability onto other human actors? And how might future models of robotic accountability require this deflection to be controlled? At stake is the potential ultimately to protect against new forms of consumer and worker harm.

This paper 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. By articulating mismatches between control and responsibility, we argue for an updated framework of accountability in human-robot systems, one that can contend with the complicated dimensions of cooperation between human and robot.

Introduction[1]

After navigating through automated menus, being misunderstood by voice recognition software, and waiting on hold for fifteen minutes, it is hard not to vent anger at the woman who answers your call to the insurance company. It is common to see an airline representative at the gate of a canceled flight be yelled at by frustrated travelers, even though he neither caused the cancelation nor possesses the power to change it. On the front lines of large, bureaucratic systems, people positioned as the external interface of a system appear at once a metonym for the company and also as gatekeepers to the company. As gatekeepers, they seem to possess a degree of agency, a capacity to take effective action, which the customer does not. But in general, we know that such individuals do not represent the whole company, and that agency is only perceived, not actuated. We know, in most cases, these individuals are not responsible for the decisions that have led up to the situation.

In instances like these, humans at the interface between customer and company are like sponges, soaking up the excess of emotions that flood the interaction but cannot be absorbed by faceless bureaucracy or an inanimate object. There may be affective ramifications for this misplaced blame, but the discerning customer or manager will know that the individual is not responsible. However, in automated or robotic systems it can be difficult 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, who or what is in control?

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.[2] A prevailing rhetoric of human-computer interaction design

[1] Support for this research has been provided by a grant from the John D. and Catherine T. MacArthur Foundation and has been conducted as part of the Intelligence & Autonomy Initiative at Data & Society Research Institute. I would like to thank Tim Hwang, Robin Sloan, danah boyd and my colleagues at Data & Society for the countless insights, discussions and suggestions that have shaped this paper.

[2] Michael Chui, James Manyika and Mehdi Miremadi. 2015. "Four Fundamentals of Workplace Automation." *McKinsey Quarterly.* November. http://www.mckinsey.com/business-functions/business-technology/our-insights/four-fundamentals-of-workplace-automation (accessed 1/2/2016); Thomas Davenport and Julie Kirby. "Beyond Automation." *Harvard Business Review*. June. https://hbr.org/2015/06/beyond-automation (accessed 1/22/2016).

suggests that keeping a "human in the loop" assures that human judgment will always be able to trump automation.[3] 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 accountability.

In a previously published case study of the history of aviation autopilot litigation, Tim Hwang and I documented a steadfast focus on human responsibility in the arenas of law and popular culture, even while human tasks in the cockpit have been increasingly replaced and structured by automation.[4] Our analysis led us to thinking about the incongruities between control and responsibility and the implications for future regulation and legal liability in intelligent systems. The dilemma, as we saw it, was that as control has become distributed across multiple actors (human and nonhuman), our social and legal conceptions of responsibility have remained generally about an individual. We developed the term *moral crumple zone* to describe the result of this ambiguity within systems of distributed control, particularly automated and autonomous systems.[5] 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—

---

[3] See M.L. Jones's WeRobot 2014 presentation for a related discussion of the ironies of automation law, Meg Leta Ambrose. "Regulating the Loop." WeRobot 2014.

[4] M. Elish and T. Hwang. 2015. "Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation. Data & Society Intelligence & Autonomy Working Paper. http://ssrn.com/abstract=2720477 (accessed 2/1/2016). Our conclusions are supported by similar work. For instance see David Mindell. 2015. *Our Robots, Ourselves*. New York: Viking.

[5] 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 is useful for specifically analyzing the types of perceptions and actions at stake in autonomous systems. Parasuraman et al. 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. Parasuraman et al. 2000. "A Model for Types and Levels of Human Interaction with Automation." *IEEE Transactions on Systems, Man and Cybernetics* (30)3.

accidentally or intentionally—that bears the brunt of the moral and legal responsibilities when the overall system malfunctions.

This paper presents a set of challenges at stake in shared control by examining how responsibility for the operation of a system may be misaligned with how control is exercised within that system. First, I present two cases in which I suggest that moral crumple zones emerge, the partial nuclear meltdown at Three Mile Island and the fatal crash of Air France flight 447. The circumstances surrounding 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 posses only limited knowledge and control. In the final section of the paper, I argue that an examination of the mismatches between control and responsibility are relevant beyond industrial contexts and that cultural perceptions of the role of humans in automated and robotic systems need to be updated in order to protect against new forms of consumer and worker harm.

The Accident at Three Mile Island

The Three Mile Island Nuclear Generating Station is a nuclear power plant on the Susquehanna River ten miles southeast of Harrisburg, Pennsylvania's capital. Three Mile Island was the eighth nuclear power plant to be built in the United States and the largest, consisting of two units, one of which is currently operating. 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. It was the first nuclear disaster in the United States, and was a major blow to the development of the civilian nuclear energy industry.[6] Although the accident occurred several decades ago, the challenges of shared control between humans and machines that contributed to the accident remain essentially unsolved.

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

[6] P. Behr, "Three Mile Island Still Haunts U.S. Nuclear Power Industry," New York Times, 27 Mar 09. http://www.nytimes.com/gwire/2009/03/27/27greenwire-three-mile-island-still-haunts-us-reactor-indu-10327.html(accessed 3/1/2016).

electrical energy. TMI is a Babock & Wilcox reactor, which consists 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 nuclie fission occurs, controlled chain reactions that split apart atoms, 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 then 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, every risk was calculated, planned for, and addressed by the putative fully automated system.[7]

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, and in the early morning of March 28, one of these filters became clogged. It would later come to light that these filters had consistently caused problems that the plant management had ignored.[8]

The shift supervisor, William Zewe, a graduate of the Navy's nuclear operations program, oversaw a team of several dozen people during his eight-hour shift at the plant. None of the positions required advanced knowledge of nuclear energy or systems design, and training courses were short and basic. Two of operators on duty on March 28, Craig

[7] Note about automation planning and design in nuclear systems of 1970s.
[8] Ford 1981: 95.

Faust and Edward Frederick, had backgrounds as enlisted personnel who operated submarine reactors for the Navy.

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 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.

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.[9] 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 to the operators about the system.

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. 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.[10]

The nuclear accident at Three Mile Island provides an example of the emergence of a moral crumple zone. Based on press releases from plant management, the governor's office, and the Nuclear Regulatory Commission (NRC), news coverage in the weeks following the accident laid unequivocal blame on the plant operators. A *Los Angeles Times* front-page headline from April 11, less than two weeks after the meltdown, stated "Nuclear Accident Blamed Primarily on Human Error."[11] Reporting on the official NRC report that was released two months later, one Associated Press headline read, "Human Error Cited in 3-Mile Accident."[12] The first paragraph stated: "Operators of the Three

---

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

[10] The findings of this report, known as the Kemeny report, particularly emphasized the role of the reigning "mindset" at the plant and how "systemic" problems were the basis for the accident. J. G. Kemeny et al., "Report of the President's Commission on the Accident at Three Mile Island," U.S. Government Printing Office, 0-303-300, October 1979. http://www.threemileisland.org/downloads/188.pdf (accessed 2/8/2016).

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

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

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 is 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."[13]

In the opening minutes of a PBS American Experience 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."[14]

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.

If the frame is expanded beyond those immediately present during the accident, these errors followed directly from other systemic errors. The workers had been directed to test these valves and document the testing in a way that cut corners and saved money and time for the plant managers.[15] 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.[16] 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

---

[13] Ibid.
[14] PBS WGBH. "Meltdown at Three Mile Island." *PBS American Experience*. Transcript. 1992. http://www.pbs.org/wgbh/amex/three/filmmore/filmcredits.html(accessed 3/1/2016).
[15] D. Ford. 1981. "A Reporter at Large: Three Mile Island." *New Yorker*. Apr 6: 49-120:111.
[16] J. Omang. "'Nuggets': A Collection of Nuclear Glitches." *Washington Post.* 10 February 1979. https://www.washingtonpost.com/archive/politics/1979/02/10/nuggets-a-collection-of-nuclear-glitches/a48dbfae-d6d8-45ef-964c-a433a5f6bdf6/(accessed 3/10/2016).

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.[17]

Furthermore, the control room design, as mentioned earlier, did not provide adequate information or feedback to allow operators to assess the state of the system. The operators made incorrect decisions because they had incorrect information. Focusing on the agency of operators misses other 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.

The Crash of Air France Flight 447

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.[18] Viewed in a different light, the circumstances of the accident provide a paradigmatic example of how human operators become moral crumple zones in complex system failures.

After an on-time departure from Rio de Janeiro, the flight proceeded for one hour and forty minutes without incident. 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. Since the late 1980s, the FAA requires the presence of two pilots in the cockpit during flight.[19] The most senior pilot and the Pilot in Command, ultimately responsible for the flight, was Captain Marc Dubois. Also flying was Pierre-Cédric Bonin and David Robert, both relatively young pilots who had spent the majority of their flight hours in

[17] Ford 1981. See also Kemeny et al. 1979.
[18] W. Langewiesche. 2014. "The Human Factor." *Vanity Fair* September. http://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash (accessed 10/4/2015).
[19] For the preceding decades, three crewmembers were required, two pilots and a flight engineer. The automation of most of the systems originally controlled by the flight engineer obviated this role.

Airbus aircraft in which pilots spend more time monitoring systems than actively controlling the aircraft.

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.[20] 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, which sounds like it might refer to an alternate universe, in fact refers to a mode in which primary control is in the hands of the pilot. When the computer and autopilot are unable to work as designed, like if a sensor reading is absent, Alternate Law is engaged. In Normal Law, the computer would override any actions that would result in an aerodynamic stall, which results from an incorrect angle of attack, the degree

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

at which the airplane wing meets the oncoming air. In Alternate Law, the pilots are essentially on their own.

To return to the accident timeline, about an hour and half into the flight, Captain DuBois left the cockpit to nap in the crew quarters and David Robert joined Bonin in the cockpit. From the transcript recovered from the black box, it seems that Bonin was anxious about a storm that he could see on the radar. As they reached the storm, they encountered ice crystals that accumulated in the airplane's pitot tubes, sensors which measure airspeed. Frozen, the pitot tubes could not transmit airspeed indications, which the autopilot requires to function. With a "cavalry charge" alarm, the plane reverted to Alternate Law and the pilots learned that the autopilot had disengaged. Soon another alarm sounded, indicating a deviation in planned altitude. Bonin, 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." Within a few more seconds, the pilots realized that the autopilot had failed because of incorrect speed indications.

At this point, the pilots should have had enough knowledge and time to fix this relatively simple problem of recovery from an aerodynamic stall. 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 recovering from the stall.

Instead of lowering the nose of the plane, Bonin pulled back on the control stick, raising the nose of the plane trying to climb. In the following minute, numerous alarms went off as Bonin frantically tried to control the plane. Likely adding to Bonin's 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 Bonin's confused state, he had brought the plane up as high as 23 degrees. Communication between Bonin and Robert had broken down, and while Robert seems to have tried to take control of the plane, the design of the Airbus controls only allow one pilot to be in control at a time, but 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. Robert was pushing forward, Bonin pushing back. Neither one aware of the actions of the other. One minute and seventeen seconds had passed since the reversion to Alternate Law.

Twenty-one seconds later, and finally responding to their summons, Captain DuBois returned to the cockpit. 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 three pilots helpless. 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 Bonin 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. According to the recovered audio, at one point Robert said to Dubois, "We completely lost control of the airplane, and we don't understand anything! We tried everything!" 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 headlined the role of pilot error in the crash. American news outlets, quoting the official French report stated, "a series of errors by pilots and a failure to react effectively to technical problems led to the crash."[21] Many of the details described above were subsumed under a narrative in which the pilots lost "cognitive control" and caused the crash.[22] A typical news report, here from CNN, explained,

---

[21] BEA. 2012. *Final Report on the Accident on 1st June 2009*. https://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf (accessed 1/17/2016).

[22] Pilot error has been a consistent catchall for explaining commercial and private aircraft accidents. See S.A. Leveen. 1982. "Cockpit Controversy: The Social Context of Automation in Modern Airlines." Ph.D. Dissertation, Department of Science and Technology Studies, Cornell University.

> 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.[23]

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 also important to consider the larger structural circumstances that in many ways primed the pilots for "the total loss of cognitive control." 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.[24] More free time does not necessarily lead to high-level judgments. In fact, pilot awareness generally decreases with increased automation.[25] Human factors research has demonstrated that skills atrophy when automation takes over.[26] While the senior pilot, DuBois 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.[27]

The problems arise not only in deskilling, but also in the kinds of interactions expected between pilots and the flight management systems. Regulators, in addition to the engineers and managers of aviation systems, have created a schizophrenic dynamic in which automation is seen as safer and superior in most instances, unless something goes wrong, at which point humans are regarded as safer and superior. Unfortunately, creating

---

[23] CNN Wire Staff. 2012. "Final Air France crash report says pilots failed to react swiftly" *CNN.com* 5 July. http://www.cnn.com/2012/07/05/world/europe/france-air-crash-report/ (accessed 1/17/2015).
[24] L. Bainbridge. 1983. "Ironies of Automation." *Automatica* 19: 775-779; R. Parasuraman and V. Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* June 39(2): 230-253.
[25] S. M. Casner and J. Schooler. 2014. "Thoughts in Flight: Automation Use and Pilots' Task-Related and Task-Unrelated Thought." *Human Factors* 56(3), 433-422;
[26] N.B. Sarter, D.D. Woods, and C.E. Billings. 1997. "Automation Surprises" in *Handbook of Human Factors & Ergonomics 2nd edition*, G. Salvendy ed. New York: Wiley.
[27] Langewiesche 2014.

this kind of role for humans, who must jump into an emergency situation at the last minute, is something humans do not do well.[28] Under these circumstances, the odds are stacked against the pilot.

Still, the rhetoric around the infallibility of automation persists. Consider 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 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.[29]

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 is 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."[30] Courts have consistently

---

[28] A.H. Roscoe. 1992. *Workload in the glass cockpit. Flight safety digest.* Alexandria, VA: Flight Safety Foundation; Earl L. Weiner. 1989. *Human factors of advanced technology ("glass cockpit") transport aircraft* (NASA Contractor Report 177528). Moffett Field, CA: NASA Ames Research Center.

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

[30] 14 CFR 91.3 The pilot (and by extension in most cases, airline) is responsible for the plane's operation whether she uses the autopilot, chooses not to use the autopilot, uses the autopilot incorrectly or acts incorrectly because the autopilot gives faulty information. J. E. Cooling and P. V. Herbers. 1983. "Considerations in Autopilot Litigation." *Journal of Air Law and Commerce* 48, 693-723: 713.

upheld this authority of the pilot as the ultimate designation of liability.[31] While control has been effectively distributed,[32] responsibility has not scaled accordingly.

Moreover, the framework of autopilot certification bounds the automatic system in a way that limits accountability to only mechanical failure. The first two sub-points of autopilot certification requirements dictate the necessary ability of the autopilot to be disengaged by the pilot. Specifically "each system must be designed so that the automatic pilot can:

1. Be quickly and positively disengaged by the pilots to prevent it from interfering with their control of the airplane; or
2. Be sufficiently overpowered by one pilot to let him control the airplane.[33]

In the most recent version of 14 CFR 23.1329 (2011), the amount of force and time to positively disengage the autopilot are specified. The following are the reasonable periods of time established for "pilot recognition between the time a malfunction is induced into the autopilot system and the beginning of pilot correct action following hands-off or unrestrained operation":

1. A three-second delay following pilot recognition of an autopilot system malfunction, through a deviation of the airplane from the intended flight path, abnormal control movements, or by a reliable failure warning system in the climb, cruise, and descent flight regimes.
2. A one-second delay following pilot recognition of an autopilot system malfunction, through a deviation of the airplane from the intended flight path, abnormal control movements, or by means of a reliable warning system, in maneuvering and approach flight regimes.

More simply stated, an autopilot must be designed to allow the pilot three seconds to correct a malfunction and still maintain safe flight when climbing, descending or

[31] For example, Air Line Pilot's Assoc., Int'l v Federal Aviation Administration, 454 F.2d 1052 (D.C. Cir. 1971). See also Cooling et. al 1983: 713-714.
[32] 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.
[33] 14 CFR 23.1329 1982

cruising. During periods of takeoff and landing the pilot must have a one-second time frame to correct the malfunction.[34]

The autopilot functions 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. Recall that 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 a 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 in the moral crumple zone.

Discussion

In a paper titled "Accountability in a Computerized Society,"[35] Helen Nissenbaum outlines four main barriers to the establishment of accountability, or 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 attitudes toward accountability. Her argument and the arguments developed by other philosophers of technology in recent years, analyze how the unique affordances of computational technologies obscure traditional paths to identify accountability.[36] As both a contribution and extension to this body of literature, the argument advanced in this paper is intended to articulate the role that social constructions of responsibility play in assigning blame, and also call attention to the potential for these constructions to shape future legal decisions around autonomous and robotic technologies. What is unique about the concept of a moral crumple zone is that it highlights how structural features of a system may

[34] I am not aware of any studies that support or indicate these time frames. Cooling & Herbers reached the same conclusion in 1983 and maintained their conclusion in 2015. J. Cooling and P.Herbers. 2015. Personal communication with author. Feb 16.
[35] H. Nissenbaum. 1996. *Science and Engineering Ethics* 2(1): 25-42.
[36] For example, DG Johnson 2006. "Computer systems: Moral entities but not moral agents." *Ethics and Information Technology* 8: 195–205; Coeckelbergh, M. 2011. "Moral Responsibility, Technology, and Experiences of the Tragic: From Kierkegaard to Offshore Engineering." Science and Engineering Ethics 18(1): 35-48.

inadvertently take advantage of human operators (and their tendency to become sponges) to fill the gaps in accountability described by Nissenbaum and others.

Still, moral crumple zones do not emerge in 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…."[37] 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.[38] Indeed, Nissenbaum uses the Therac-25 accidents as an example of the "the problem of many hands,"[39] 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.[40]

In the context of this paper, the question that concerns us is: why do perceptions of accountability stop at the computer sometimes, but at other times create moral crumple

[37] Associated Press. 1986. "Fatal Radiation Dose in Therapy Attributed to Computer Mistake." *New York Times* Jun 21, pg. 50.
[38] N. Leveson and Clark Turner have written an extensive report on the accidents and provide analysis on lessons learned for future safety-critical software system development. N. G. Leveson and C.S. Turner. 1993. "An Investigation of the Therac-25 Accidents." *IEEE* July:18-41.
[39] Nissenbaum borrows the phrase from D. Thompson. 1980. "Moral responsibility and public officials: The problem of many hands." *American Political Science Review* 74(4): 905–916.
[40] Nissenbaum: 8.

zones in which humans are caught?[41] Perhaps it is that in the accidents involving Therac-25, the mistakes and oversights of the manufacturer were so egregious that there could be no mistaking that it was the software that was in control of administering the doses of radiation.[42] More to the point, the defect was *recognizably* egregious because it violated an existing standard or certification.

Thus far, the discussion has focused not on legal liability but rather on cultural perceptions of blame and responsibility, particularly in an American context. I would now like to turn briefly to the relationship between perceptions of accountability and legal liability. The cases of Three Mile Island and Air France are large and industrial systems, not personal or widely available commercial technologies. The scale and nature of these systems involves configurations of legal liability in which individuals are uniquely protected by unions or employers.[43] In these specific cases, legal liability was taken on by the large corporations for whom the individuals worked and who were in a position to make settlements outside of court.

As an anthropologist, my interest lies in the cultural perceptions of responsibility in automated and robotic systems, and the extent to which these perceptions permeate formal frameworks of accountability, from regulation to litigation to performance evaluations. Especially in the context of emerging technologies, social norms and expectations play a significant role in the legal integration of a technology into existing

---

[41] Another instance in which a moral crumple zone was structurally prevented from emerging was the incident involving the U.S.S. Vincennes, which shot down and destroyed an Iranian civilian passenger plane, killing all 290 people on board, in 1988. According to reports at the time and the official investigation conducted by the U.S. Navy, the seamen on board the warship, which was equipped with the most sophisticated radar at the time, mistook the plane the plane for a fighter plane. They believed the threat existed because this was the information presented to them by the Aegis radar system. The U.S.S. Vincennes is often used as a case in which the dangers of trust in automation are illustrated, and rightly so. However, the structures of the military, and also the rights that protect military action, prevented the discussion of fault and blame from focusing solely on the humans in the system. The evasion of responsibility for the accident in the arena of global politics is worthy to be discussed at length, but cannot be addressed within this scope of this paper.

[42] This might be the case in a recent accident involving a Google driverless car, the first purported accident to be caused by a Google driverless car. I will revisit accountability in driverless cars in the final section of this paper.

[43] In our work on liability and autopilot litigation, Hwang and I conjectured that one explanation might derive from the unequal power dynamics between airline crew and large airlines and manufacturers. Pilot error is the most convenient explanation for all parties except the pilot. This might also be true for the operators at a nuclear power plant.

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.[44]

Framing cultural perceptions of accountability in the context of moral crumple zones can provide a means to think about how risk is or should be distributed in systems. With regard to autonomous and robotic technologies, the regulations, laws, and norms are still in formation, and may be particularly susceptive to the bias that moral crumple zones present. 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.

Preparing for Moral Crumple Zones

While it is possible that the concept of a moral crumple zone only holds in the case of industrial systems, I would argue that the concept is useful in thinking through the regulatory and liability implications of all automated and autonomous systems. To demonstrate its utility as a framing concept or provocative wrench, I present first a hypothetical scenario and then discuss near term instances in the transportation sphere where we might see moral crumple zones emerge.

Education

One arena in which we are likely to see the implementation of intelligent and automated systems soon may be in educational settings.[45] Consider a hypothetical scenario that involves a teacher working with an automated, and personalized, virtual teacher. The teacher's primary role becomes monitoring a large class of students as they

---

[44] See for instance, A. M. Froomkin. 1995. *The Metaphor is the Key: Cryptography, the Clipper Chip, and the Constitution*, U. PA. L. REV. 709, 861-62, and R. Calo. 2016. *Robots in American Law*. We Robot 2016 presentation. Legal Studies Research Paper No. 2016-04.

[45] N. Singer. 2015. "Silicon Valley Turns Its Eye to Education." *New York Times*. Jan 11. http://www.nytimes.com/2015/01/12/technology/silicon-valley-turns-its-eye-to-education.html?_r=0 (accessed 3/1/2016); A. Sneed. 2012. "Coming Soon to a Kindergarten Classroom: Robot Teachers." *Slate.com* http://www.slate.com/blogs/future_tense/2012/08/06/robots_may_become_elementary_school_teachers_in_the_future_.html (accessed 3/1/2016).

interact individually with a program on a computer tablet. In this scenario, who will be responsible for the success or failure of a student's progress? Who should be responsible? The teacher? The software designer? The student? It is easy to imagine how the teacher might be caught in a moral crumple zone: a student's parent might come to the school and blame the teacher if progress is not being made, or a teacher might be evaluated by her supervisor based on the overall performance of "her" class, even though she is no longer in primary control of the lesson plans and day-to-day teaching. As she monitors more, and teaches less, her teaching skills might atrophy, leaving her even less likely to succeed when she is called upon to teach. Caught in the moral crumple zone, we can also see how quickly the nature of her job may change, even though her perceived responsibility might remain.

Transportation

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. Do self-driving cars create moral crumple zones? Possibly.

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.[46] Already Elon Musk, referring to a new release of Tesla Autosteer software, has emphasized,

> 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.[47]

[46] M. Ramsey. 2015. "Who's Responsible When a Driverless Car Crashes? Tesla's Got an Idea." *Wall Street Journal* 13 May. http://www.wsj.com/articles/tesla-electric-cars-soon-to-sport-autopilot-functions-such-as-passing-other-vehicles-1431532720 (accessed 8/2/2015).
[47] B. Sorokanich. 2015. "Tesla Autopilot First Ride." *Road & Track*. October 14. http://www.roadandtrack.com/new-cars/car-technology/news/a27044/tesla-autopilot-first-ride-almost-as-good-as-a-new-york-driver/ (accessed 2/11/2016).

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.

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.[48]

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. Another example might be seen in the current discourses around driverless car accidents. In the summer of 2015, Google made public the accident record of its self-driving car tests. The announcement and subsequent press coverage declared that none of the accidents had been caused by the Google car; all were the fault of human drivers. As a safety precaution, a human driver had always been present during testing, prepared to take over if anything went wrong, and in fact, one of the times a Google car was in an accident was when it was being driven entirely by a human in a parking lot.

Still, there was a surprising pattern of rear-end accidents, ten out of twelve. Perhaps these kinds of accidents are the most common on the stop-and-go streets of Palo Alto.[49] 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

---

[48] J. Markoff. 2016. "Google Car Exposes Regulatory Divide on Computers as Drivers." *New York Times*. Feb 10. 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).

[49] As Schoettle and Sivak point out in a comparative study of accident rates between driverless and traditional cars that the datasets of driverless cars are much smaller, and in limited settings, compared to traditional cars. This is a weakness of any current comparative studies. B. Schoettle 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. http://www.umich.edu/~umtriswt/PDF/UMTRI-2015-34_Abstract_English.pdf (accessed 10/12/2015).

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? 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. In this vein, and as a final point, I would like to draw attention to certification conceptualizations and processes as a productive area of future research.

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 certainly come up in regards to deep learning technologies, and other emergent forms of artificial intelligence.[50] How do you certify what is theoretically an unbounded system? While I have no answers, I believe that a productive area of inquiry will be an examination of best practices in certifications for intelligent and autonomous systems. How can systems be certified as safe and effective while not creating moral crumple zones?

Conclusion

This paper 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

[50] This is next major step for the Google self-driving car project: Chris Umson. Letter. NHTSA http://isearch.nhtsa.gov/files/Google%20--%20compiled%20response%20to%2012%20Nov%20%2015%20interp%20request%20--%204%20Feb%2016%20final.htm (accessed 3/10/2016)

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 paper, 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. Because moral crumple zones arise in the context of distributed control, I argued that moral crumple zones have the potential to exist in consumer systems. I also explored how traditional modes of technology certification may reify the potential to create moral crumple zones, and suggest that a reexamination of certification paradigms may be a productive avenue of future research. This paper 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.