On the Moral Hazard of Autonomy

A. Terry Morris, Jeffrey M. Maddalon, Paul S. Miner
Safety-Critical Avionics Systems Branch
NASA Langley Research Center
Hampton, Virginia, United States
Email: {allan.t.morris, j.m.maddalon, p.s.miner}@nasa.gov

Abstract*—This paper describes the concept of moral hazard as applied to technologies that incorporate automation and autonomy. Moral hazard is said to exist when a party to a transaction feels more comfortable taking undue risks because another party will bear the costs if things go badly. As opposed to regular physical hazards, a moral hazard comes from within a person. In this paper, we reveal two categories of moral hazards related to autonomy. The first category of moral hazard occurs when the owner of the autonomy introduces an autonomous system without accepting the full responsibility for improper operation thereby shifting the risks from one party to another party. This category of moral hazard is similar to moral hazards experienced in other industries and can often be addressed through appropriate policy and establishing liability for irresponsible behavior. The issue becomes more complicated in cases where the operator of the autonomy may not have a full understanding of the system behavior. In the second category of moral hazard, risks are shifted from people to autonomy. In this category, the humans in proximity to the autonomous system begin to trust its behavior. Their behavior may change in that they may believe they are more insulated from harm and subsequently exhibit more risky behavior towards increasingly autonomous technologies. Mitigating this type of moral hazard may require the autonomy to possess certain design features to discourage this type of harmful human behavior so that humans do not suffer needlessly in their interactions with autonomous systems by placing inappropriate trust where that trust is neither warranted nor deserved.

***Keywords*—moral hazard, hazards, autonomy, artificial intelligence, human machine teaming, automation complacency**

I. Introduction

For decades, engineers have used the techniques from System Safety Analysis across a wide variety of fields to develop safety-critical systems. For the most part, these techniques have proven effective at reducing harm, thus refuting the premise of some predictions [1]. However, in the last few years, a number of factors have come together that portend a new type of system, those based on Artificial Intelligence (AI) that will exhibit a new type of hazard that has less in common with the traditional hazards discovered through System Safety Analysis and more in common with hazards that appear in, heretofore, primarily human realms of finance and insurance. We term these hazards moral hazards, due to their similarity to the related hazards in these other fields. Through identification of these hazards, and proper mitigation of them, the general public will trust the safety and performance of these autonomous systems in a similar manner to how the public trusts many of the complex safety-critical systems that they already use. For the purposes of this paper, we use the term autonomy to mean a system that uses AI technologies to make decisions where generally humans cannot intervene in the decision-making process. Furthermore, we define a safety-critical autonomous system to be an autonomous system whose decisions have the potential to cause significant harm to human life. Given the increased rate of adoption of these new AI technologies, some believe that the maturity of fully trusted autonomous systems is a few months away while others believe it will not occur for several decades [2].

We recognize that this paper is speculative in nature and we are extrapolating from some recent trends in the industry. Depending on how these technologies develop, the ultimate relevance of this paper is uncertain. We hope readers will indulge this speculation, since our motive is to set the foundations for a discussion of how safety engineers are to address the unique challenges as humans interact with autonomy. As such, this paper is not intended to provide best-practices for using particular algorithms, middleware frameworks, data architectures, etc. As important as those issues are, they are best addressed by communities of specialists in those techniques.

II. Relevant Advances in AI and Autonomy

Since the dawn of the computer age, some of the luminary minds of computer science have dreamed of computers that provide human-like thinking abilities such as Vannevar Bush [3] and Allan Turing [4], which is not to neglect the science fiction writers who came earlier, but perhaps with a less practical path to realization [5]. However, in the last few years, a number of advances have been made that when coupled perhaps will fulfill at least some of the dreams of these earlier thinkers. Specifically, the emergence of massive floating-point computer power in the form of Graphics Processing Units (GPU), large, fast memory storage in the form of Double Data Rate 3 Synchronous Dynamic Random-Access Memory (DDR3 SDRAM), and large data sets of images available at various Internet photo archives have allowed machine learning algorithms to sometimes solve the "tell me what is in this picture" problem in limited degrees under specific circumstances and to a fixed level of abstraction. Apocryphally, in 1966, Marvin Minsky gave this project to a first-year graduate student to solve over the summer [6]. This perception problem has been one of the great unsolved problems within AI. In addition to its challenging nature, this problem is of massive importance to many domains, and plays a central role in the many safety-critical autonomous systems for transport applications.

Table 1. Machines outperforming humans in complex games and tasks [7].

Date	Game/Task	Outcome
2011	Jeopardy	IBM’s Watson beats two former champions to win Jeopardy
2014	Facial recognition	Facebook’s DeepFace AI facial recognition algorithm achieves and accuracy rate of 97%, rivaling the rate of humans
2015	Go	Google DeepMind’s AlphaGo defeats Go champions in Korea and Europe
2016	Speech recognition	Microsoft speech recognition AI can transcribe audio with fewer mistakes than humans
2017	Poker	Libratus, and AI bot, defeats four of the world’s leading poker players in a 20-day tournament
2017	Visual intelligence test	An AI system developed by Northwestern University is able to beat 75% of Americans at a visual intelligence test
2018	Reading Comprehension test	Alibaba’s AI outscores humans in a Stanford University reading comprehension test

With the long history of AI, one may question whether this claimed revolution in computing is truly near at hand. It may be helpful to examine some recent, impressive outcomes of these machine decision-making systems. In the past decade, these machines have incorporated technologies that have surpassed humans in complex games (Chess, Go, Jeopardy, poker, etc.) or have reached levels on par with humans (see Table 1) [7]. With the pace of these technological advancements, what company would not want to incorporate these advancements to be first to market or to gain market share. The incentives to incorporate these AI technologies into present-day systems is enormous. These technologies, however, have downsides that need to be identified, mitigated and managed.

As impressive as these accomplishments are, they do not quite trigger the potential massive implications that AI can have on society. The Device for the Autonomous Bootstrapping of Unified Sentience or DABUS was the inventor for a patent. On April 22, 2020, the United States Patent Office (USPTO) issued a decision that US patent law is limited to natural persons, rejecting an application for an invention by an artificial intelligent machine [8]. The petitioner (a human) asserted that the invention was developed by a creativity machine named DABUS, which was “trained with general information in the field of endeavor to independently create the invention” [9]. The USPTO rejected the petitioner’s argument, stating that the “granting of a patent under 35 U.S.C. § 151 for an invention that covers a machine does not mean that the patent statutes provide for that machine to be listed as an inventor in another patent application any more than a patent for a camera allows the camera to hold a copyright” [9]. The USPTO took the position that inventorship in the US is limited to natural persons. The particular wording used in relevant statutes provided the basis for this reasoning and the reasoning was supplemented by various federal circuit decisions in other nonpatent contexts. The USPTO also found judicial reasoning to support the notion that conception of an invention relies upon mental processes, not simply any act of creation. In like manner, the petitioner brought his request to the European Patent Office (EPO) and the UK Patent Office (UKIPO) to receive similar outcomes [8]. Notably, the UKIPO determined that nonhumans can create intellectual property but lack the requisite personality to claim those rights.

The ramifications of this decision are simple – humans may not be the only entities that can create. They are, however, the only entities capable of claiming the rights associated with invention. However, what if this decision had gone the other way? If DABUS had been granted rights as an inventor, what expectations should society have on DABUS to act as an inventor? Setting aside the political or philosophical implications, what steps should engineers take to ensure that systems such as DABUS meet society’s expectations? For instance, if another inventor claims that DABUS “stole” his work, DABUS may be required to be able to reveal how it arrived at its invention, as if it was testifying in court.

III. Artificial intelligence applied to safety-critical systems

AI is advancing at an unprecedented rate due to affordable computational power and a concentrated focus on the field by tech giants who have computer engineers actively discussing age-old philosophical problems [10]. The pace of technological advancements is forcing governments, companies and society to ask questions related to utilitarianism, consequentialism and fairness. Technologists are now grappling with philosophical concerns and moral dilemmas with no clear answers. With the advent of smart machines with learning capabilities powered by artificial intelligence, we need to find practical solutions for dealing with these technologies [10]. If we allow complex, decision-making autonomous machines to make safety-critical decisions, then we must be ready to deal with the positive and negative consequences of that choice.

Based on society today, the harm to a human (through faulty decision logic, bad software or sensor malfunctions) is a moral decision with legal ramifications. Some philosophers state that only moral agents can be the bearers of moral obligations, duties, and responsibilities [11]. Only autonomous moral agents (mature humans) are ascribed full moral responsibility for their behavior. How can a large complex system with autonomous technologies making safety-critical decisions function adequately in human society if the autonomous system is not held responsible for its behaviors? As of today, our legal system does not provide moral rights to a machine. The designers, manufacturers and operators of systems with Increasingly Autonomous (IA) technologies are held liable for inadequate behavior that may contribute to the loss of human life. This, however, may change in the future with the rate of technological advancement and a subsequent changing of the applicable laws.

Today, we design safety-critical systems and ensure that they are safe and secure. We verify that the system does not readily pose hazards or threats to human beings to an acceptable level of risk [12]. Risk, it appears, is the means by which we develop and design systems and the approving rationale why we validate, verify and accept such systems. In the spectrum of risk analyses, we analyze the physical hazards of the system, the environment and other phenomenological hazards and attempt to mitigate these potential hazards by various approaches (design for minimum risk, dissimilar redundancy, common cause analysis, etc.). We also analyze the security risks and the safety risks to society.

In the area of unmanned aircraft systems (UAS), the plethora of technologies envisioned for advanced air mobility (AAM) is astounding [13]. When operators utilize UAS for flight near people, how can they ensure that their vehicle is safe? Additionally, how can an authority develop objective certification criteria when the UAS designer cannot fully explain how the embedded technology makes decisions? NASA and other agencies are currently working with the FAA to draft candidate certification approaches for UAS that incorporate IA systems [14]. This will undoubtedly take some time. Until there is a bonafide certification program for UAS in the National Airspace System (NAS), each stakeholder assumes a myriad number of risks to themselves and to the public.

In analyzing system risks, we have noticed that there are other hazards that may manifest when humans interact with autonomous systems in safety-critical domains. These additional risks are called the moral hazard of autonomy and they represent the increased risk a person takes when someone else bears the costs if things go badly in a transaction involving autonomous systems. This paper will describe the moral hazard of autonomy and will describe mitigation approaches to prevent or to reduce the likelihood of these risks.

IV. HAZARDS AND MORAL HAZARDS

A. System Safety Hazards

The concept of system safety involves the utilization of risk-based strategies to identify, assess, eliminate and mitigate the various hazards of that system. These system safety analyses are often performed throughout the life cycle of the system from systems requirements to system verification to system disposal. The system in question is defined as a group of interdependent and interrelated elements working together to achieve a common objective. This definition emphasizes the importance of the interactions between system components and the environment (including human interactions). The aim of the system safety concept is to model, analyze, and understand the hazards. This knowledge is aimed at either the elimination or, at least, mitigation of all relevant hazards of the system so that the entire system can achieve an acceptable level of safety. Some system safety analysis techniques include integrated hazard analysis [15], functional hazard assessment, fault tree analysis, common cause failure analysis, sneak circuit analysis and failure modes, effects and criticality analysis to name a few. For countries that demonstrate safety through the process of certification, each safety critical industry has a set of governing documents and guidelines used to certify the safety of each system. For instance, aircraft certification is governed by ARP 4754 and ARP 4761 in conjunction with other aviation standards such as DO-178C and DO-254. Hazard analysis, in the context of ARP 4761, attempts to identify the complete set of system level functions, to determine how each function can fail, to ascertain the consequences of each functional failure and subsequently to design mitigations to improve the situation.

B. What is Moral Hazard?

According to Kaplan's Glossary of Insurance Terms, a hazard is a specific situation that increases the probability of the occurrence of loss arising from a peril [16]. Kaplan also explains that a moral hazard is a condition of morals or habits that increases the probability of loss from a peril. As opposed to regular physical hazards, a moral hazard indicates that the hazard comes from within a person. This implies that a moral hazard can be created based on what a person believes is the right way to act in a given situation. Subsequently, a habit can also create a moral hazard in part because we humans build habits based on what we perceive to be acceptable ways of behaving. In short, humans may begin to behave differently the more we interact with new situations or new technologies. Moral hazard exists when a party to a transaction feels more comfortable taking undue risks because another party will bear the costs if things go badly. Although this concept has traditionally been applied to finance and insurance industries—typically through people's behavior when they are insured against losses [17]; this concept can also be applied to technical domains. Moral hazard can occur under conditions called adverse selection or principal-agent scenarios.

In adverse selection scenarios, one party makes a decision based on limited or incorrect information, which leads to an undesirable result. Adverse Selection represents behaviors that occur before a contract is signed. These behaviors are usually due to hidden information or "information asymmetry" and are generally problematic in contract negotiations. In principal-agent scenarios, the principal generally delegates authority to the agent to act or make decisions on their behalf. Conflicts of interest arise between the principal and the agent because people (agents), though contracted to behave one way, tend to act in their own best self-interest. When the agent makes decisions and/or takes actions on behalf of the principal and the self-interests of the agent emerge, the environment is ripe for moral hazard. The principal, in order to manage the contractual process to some degree, will have to spend money on monitoring and providing incentives to the agent because it is generally impossible for the principal at zero cost to ensure that the agent will make optimal decisions from the principal's viewpoint [18]. A fundamental approach to solving principal-agent conflicts is to align the incentives between the principal and the agent appropriately. These incentives can be provided or augmented into the contract. Moral hazard, in this context, usually occurs after a contract is agreed and executed [19].

C. How is Moral Hazard Mitigated?

For situations of adverse selection, moral hazard can be mitigated with the use of background checks, references, testing, certifications, and acquiring more information. Information gathering helps to balance the inequality associated with information asymmetry. After the information is gathered, the terms of the contract should be adjusted according to the information acquired.

For situations of principal-agents, moral hazard can be mitigated by monitoring, that is, consulting experts, dealing with those who are reputable, establishing regulation, ensuring warranties and guarantees are in place and instituting punishments for bad behavior (copayments, deductibles, etc.). It can also be mitigated through the use of incentives or by the use of contracts and/or collateral.

D. Moral Hazard in Different Industries

Not only is moral hazard manifest in economics, it also exists in the areas of insurance (home and automobile), the financial system and the health care industry. Mitigations for moral hazards tend to place more responsibility on the human agent or the human agent by proxy (regulations).

1) Moral Hazard in the Automobile Insurance Industry: In the automobile and home insurance industries, moral hazard occurs when there is no deductible. In this scenario, humans would have no incentive to avoid minor accidents (scratches and backing into poles). People would be much more likely to take risks that could lead to minor car damage knowing that the damage is fully covered [20]. Moral hazard, in this context, is mitigated with insurance policies that requires the policyholder to pay deductibles and copays. The reasoning behind this follows. If the policyholder has to provide additional resources for minor accidents, then they are more likely to remain vigilant in order to avoid minor accidents. The mitigation here serves to align the incentives of the policyholder (the agent) to those of the insurance provider (the principal).

2) Moral Hazard in the Financial Industry: In the financial crisis of 2007-2008, the US experienced moral hazard in the financial industry. Moral hazard occurred when the government was forced to bail out "too big to fail" banks to avoid catastrophic consequences for the entire economy. In this scenario, bankers pressured politicians to deregulate banking in 1999 with the Gramm-Leach-Bliley Act (which repealed the Glass-Steagall Act of 1932). After deregulation legislation passed, bankers were then provided government-backed insurance against their losses, which gave Wall Street incentive to take more risks with Main Street's money. As stated in the Harvard Business Review in 2009, it was worrying "...because the moral hazard imposed on the system in recent months is truly mind-boggling in scale and scope. Across the globe the banks and insurers whose errors of judgment created the bubbles have been bailed out without hesitation, at minimal cost to them but at significant potential costs to taxpayers" [21]. This moral hazard could have been mitigated by making sure that those who make decisions about how to invest other people's money face commensurate punishments [22] if they make bad investments due to their own errors of judgment [20].

3) Moral Hazard in the Health Insurance Industry: In the health care industry, moral hazard occurs if people only sign up for health care when they are sick. In this case, the health care system would be too expensive to sustain. Because everyone pays into the system, those who are healthy help to provide financial resources to those who are sick by way of insurance premiums. Moral hazard, in the context of US health care, is mitigated by copays for office visits and deductibles for health care services [20]. This ensures that those who are chronically sick (due to their own decisions) don't increase health care costs for those who are healthwise responsible. This hazard either does not exist or is managed very differently in countries with universal health care.

E. The Moral Hazard of Autonomy

As in the previous industries, the technology industry may also experience moral hazard. This is termed the moral hazard of autonomy and is somewhat different from other moral hazards in that this type incorporates machines who can make decisions that influence safety-critical contexts. Before the rise of machines that could make complex, safety-critical decisions, only humans were placed with the responsibility of critical decision-making. Even in today's complex aircraft, the human pilot is still held responsible for the lives on the plane. If safety-critical decisions can now be delegated to autonomous machines capable of adaptive behavior, who is held responsible when the machine makes decisions that harm human life? Will the machine be required to explain its decision reasoning, as if it were under oath? Can the machine be punished and what does this punishment look like? Or, is only the machine's designer held responsible for its decisions?

The answer to these questions will ultimately be resolved through legal and political action and thus are not in scope of this paper. The point is that the consequences of misbehaving autonomy can be severe, wide-ranging and unexpected. This paper examines the issues of moral hazard as it relates to highly autonomous systems specifically when human lives are at risk. In this context, autonomy refers to the operation of a system with minimal human oversight where the system can perform complex, safety-critical actions. This type of autonomy is rapidly emerging in ground vehicles, aircraft, ships, submarines, spacecraft, and health care systems. In this paper, we divide moral hazard into two types: when risks are shifted from one party to another and when risks are shifted from people to autonomy.

1) Type 1 – Risks are Shifted from one Party to Another: The first category of moral hazard occurs when the owner of the autonomy introduces an autonomous system without accepting the full responsibility for improper operation. The issue becomes more complicated in cases where the operator of the autonomy may not have a full understanding of the system behavior. This situation is associated with adverse selection in that the owner of the autonomy has information asymmetry with the operator. The owner knows critical information about safety-related deficits in their autonomy but does not share it with the purchaser or operator thereby shifting the risk (also generally known as deception by omission). In some cases, the owner of the autonomy may not fully understand the behavior of the autonomy, which leads him/her toward ignorance. Systems that employ machine learning often make it impossible to determine, a priori, what proper system behavior is, or how misbehaviors may manifest. In such cases, the sharing of responsibility between the operator of the autonomous system and the developer becomes a critical issue. This moral hazard can often be addressed through appropriate policy and establishing liability for irresponsible behavior.

2) Type 2 – Risks are Shifted from People to Autonomy: We also identify a more subtle category of moral hazard where risks are shifted from people to autonomy. In this category, the humans in proximity to the autonomous system begin to trust its behavior (perhaps more than they should). If the system regularly responds in a manner to keep people safe, then people's behavior may change in that they may believe they are more insulated from harm with increased invulnerability and subsequently exhibit more risky behavior toward increasingly autonomous technologies. Historically, humans have a bidirectional relationship with technology. As we increase our trust and reliance on these technologies, the technologies slowly induce changes in human behavior that at times may (imperceptibly) be counterproductive and counterintuitive ... perhaps bordering on paradoxical [29]. This situation introduces questions regarding the correct (or perhaps acceptable) behavior of the autonomous system. Awareness and knowledge, in these cases, are protective mitigations. Humans need to understand the roles and functions of both the autonomous system and the human's interactions with the system so that humans don't endow the autonomous system with capabilities that do not exist in the system specifications.

V. EXAMPLES OF THE MORAL HAZARD OF AUTONOMY

A. Type 1 – Risks are Shifted from one Party to Another

Compared to moral hazards in other industries, type one moral hazards for autonomy are similar in that a person, company or institution provides a service or product with known or partially known defects (deficient workmanship, inadequate safety analyses, faulty software, etc.) and does not reveal all the known defects to the purchaser or operator (shifts the risk from the dealership to the purchaser). This asymmetric information places the purchaser or operator at a distinct disadvantage. This is similar to the way the proverbial "used car salesman" glamorizes all the positive attributes of the used car giving assurances of its positive health while failing to reveal known discrepancies of the vehicle. The naive purchaser who believes the used car salesmen will likely end up with a "lemon," a vehicle that breaks down after the seven day warranty expires. The used car salesman gave assurances about the used vehicle without disclosing the discrepancies leaving the purchaser to bear the cost of the hidden deficiencies (risks). The risks, in this manner, get shifted from one party to another.

To be clear, we need to distinguish between two types of transactions (contracts). The first type occurs when one party (the seller) knowingly provides guarantees or implied guarantees of safety, yet is aware of deficiencies in safety and engineering, but does not reveal this information to the receiving party (purchaser). The seller, in this scenario, is shifting the risk to the purchaser under the guise of safety (insured against loss). In the second type of transaction, a seller provides a product "as is." In this transaction, the seller communicates clearly about all product strengths and deficiencies of which they are aware but claims up front that they will not be held responsible for service or product failures. The purchaser may be well aware, partially aware or completely unaware of all known defects (information asymmetry does not exist) and the seller has provided no assurance against loss. The purchaser, in this case, assumes all responsibility for the product or service. This second scenario is not moral hazard.

1) Example – New Drone Purchased for Use in a Parade: Billy is in high school and participates in the marching band. He lives in a country that offers student permits to fly unmanned drones over public air space and Billy has this permit. The country does have laws expecting each citizen to be responsible for harm inflicted on other citizens. Billy's marching band just received high honors for a performance at the football game a few months ago and was selected to march in the local parade. Since Billy routinely participates in the unmanned aircraft challenge at his high school, he decided to purchase an autonomous drone to capture the footage of his participation in the local parade. Eager to sell one particular drone, the seller oversold the capabilities of this particular drone by effectively guaranteeing its safety and without expressing to Billy that the loss of navigation and/or communication for this drone was a routine problem. Billy purchases this drone believing the words of the seller. During the day of the parade, Billy programs his drone to automatically lift off, autonomously fly above the parade taking pictures using known GPS coordinates and then land safely. Unfortunately, the drone experienced a malfunction (loss of navigation and communication) during autonomous flight and landed on the head of Melissa, the flute player in the band. Melissa suffered some lacerations and was taken to the hospital. During the subsequent investigation, Billy declared that he was not aware of the navigation or communications issues with the drone.

In this example, Billy had flight experience with other drones at school. His experience or hubris or belief in the drone seller led him to feel comfortable that the particular drone he purchased would work correctly and safely. Unfortunately, the seller had prior information about the detriments of this particular drone and did not disclose it (information asymmetry). In this situation, Billy, as a licensed operator, took an unnecessary risk by shifting the burden of public safety onto the seller. Billy should have asked pertinent verification and validation questions about the drone or sought out other individuals with experience with this particular drone. Also, Billy could have thought about aligning the incentives between him and the seller (as agent-principle partners) so that the seller would not be motivated to lie. At a minimum, Billy should have inquired and asked who would be held responsible if the drone damaged someone through defects in workmanship. In the end, Melissa suffered the consequences of a physical collision hazard with a drone. Billy, on the other hand, suffered a moral hazard that happened to incorporate autonomy. As stated earlier, this moral hazard is very similar to moral hazards found in other industries and matters little whether autonomy is associated with it or not. It is the same hazard found in any principle-agent pair where asymmetric information is used to shift risk due to a defective product with undisclosed risk.

B. Type 2 – Risks are Shifted from People to Autonomy

In the second category of moral hazard, risks are shifted (consciously or unconsciously) from people to automation/autonomy. In this category, the humans in proximity to the autonomous system begin to trust its behavior (perhaps more than they should). Their behavior may change in that they may believe they are more insulated from harm with increased invulnerability and subsequently exhibit more risky behavior toward increasingly autonomous technologies. Humans need education, training and understanding of increasingly autonomous systems so that humans do not suffer needlessly in their interactions with these systems. This moral hazard comes from within the human and manifests itself by placing inappropriate trust in the autonomous system.

1) Example 1 – Pedestrian Crossing Street with Numerous Autonomous Vehicles:
For example, if a pedestrian walks across the street in a busy urban area like New York City with normal human beings driving the automobiles, then the pedestrian takes the risk that some of the human drivers may be distracted and may not see them. Consequently, the pedestrian was struck by a vehicle due to its distracted human driver. In this example, the pedestrian is aware of the dangers of distracted human drivers in automobiles and adjusts their behavior accordingly. The pedestrian, in this context, is more careful and will work to reduce their risk-taking behavior [18]. However, in a different example, there is a fleet of autonomous vehicles that are programmed to avoid obstacles and to minimize casualties. In this example, a pedestrian may cross the same street while the autonomous vehicles are in motion. The question arises, will these autonomous vehicles stop or will they swerve to prevent colliding with the pedestrian? If the answer is yes, then the pedestrian may increase their risk-taking behavior because they trust the technology. In other words, they have delegated (consciously or unconsciously) the responsibility for their safety to the autonomous vehicle (this includes the designers, the architects, the manufacturers, the autonomy and the regulators). Thus, the pedestrian may feel increased invulnerability [23]. During the next encounter with other autonomous vehicles, the pedestrian may ramp up their risk-taking behavior by jumping in front of autonomous vehicles just to test the collision-avoiding technologies in new and novel ways. This is an example of the moral hazard of autonomy because the hazard comes from within the human as the human interacts with automation and/or autonomy.

2) Example 2 – Human Driver Interaction with Risk Averse Waymo Autonomous Vehicle:
Some versions of Waymo's ride-sharing, self-driving cars used in Chandler, Arizona had very risk averse driving styles. Or, at least, very risk averse from the perspective of some other human drivers on the road. Apparently, out of an abundance of caution, the Waymo vehicle would stop abruptly in situations where no human driver would expect, causing the driver following the Waymo vehicle to suddenly apply brakes [24]. Some accidents where a human driver rear-ended a Waymo vehicle have been reported. Recognizing that the system was still under development, Waymo restricted the operations of the self-driving vehicle to a relatively small area of the city and avoided high-traffic situations with benign environmental conditions (no rain or dust storms). From the anecdotes in Reference [24], the situations that Waymo's van has had trouble deciding a course of action are similar to situations that are hard for human drivers, such as merging into a busy stream of traffic, unprotected left-hand turns, atypical stopped traffic on roadways (such as lines of cars pulling into a shopping center), drivers and pedestrians violating traffic laws, and people initially acting as a group and then breaking into individual behavior (that is, the problem of tracking a single entity switching to suddenly tracking multiple entities). What is atypical is the reaction to these situations by Waymo's van: abrupt stops and extra-long waits for the situation to clear seem to be the largest complaints from other drivers [24].

Waymo is actively working on improving their systems and more recent versions may not exhibit these particular issues in this manner; so, we must be careful to draw the correct lessons from this anecdotal account. Bearing this in mind, one story points out an interesting social consequence to the perceived inadequate performance by the autonomous system: one driver reports illegally driving around the Waymo vehicle when it waits too long at an intersection [24]. If this behavior is even remotely typical, then it points to a moral hazard induced by the autonomous system. Like all moral hazards this relates to a hazard that arises from human behavior in reaction to the system. In this situation, the human driver is taking an illegal and unsafe action and, thus, the human is shifting the burden to maintain safety to the autonomy. The human driver is confident in doing this because the autonomous system has a demonstrated record of being overly risk averse. In taking this course of action, the human driver is assuming that the autonomy's aversion to risk operates in a similar manner to a human's aversion to risk. It remains to be seen if this is valid.

It is perhaps natural to assume that a more risk averse autonomous system is a safer system. However, safety, in our formulation, comes from proper mitigation of all hazards. As a system designer, type 2 hazards (shifting risk from people to the autonomy) must be identified and mitigated. With regard to identification, in choosing to design the system to be risk averse, the system designers must also evaluate what would happen if the vehicle is too risk averse. As with any hazard, different mitigations can be applied to this hazard. The list below is not intended to be complete, and not even practical; rather, the intent of providing this list is merely to indicate that there are some available options to mitigate this type of moral hazard. Here are some mitigations:

- Waymo implemented a system for the public to provide comments on their self-driving cars. This is probably a useful way for the public to provide input. This information would need to be analyzed to determine where public expectation is not being met, and therefore, where a potential moral hazard could exist. It is less clear what kind of information should be asked for and what kind of follow-up (if any) should be provided.
- Warning signs could be placed on the vehicle, indicating that this system is not fully validated. Such a sign may cause a human driver to be less willing to push the limits of the system. However, there are a few caveats. First, in general, people tend to ignore signs if they see them a lot. Also, given that Waymo is attempting to collect data about how drivers react to these vehicles, signs may interfere with data collection.
- The car could be designed to be less risk averse, and if so, the humans may be less willing to assume that they can shift the risk to it.
- The car could be designed to take unexpected, but safe, actions periodically. If so, then other drivers may be more aware of their assumption that they "know" how the autonomous car will behave.
- Another mitigation, which would not be provided by the system designers, but rather by larger society, is an assumption that autonomous cars are safe. If this assumption were made by society, then if there were ever an accident between a human driver and an autonomous car, the human driver would be blamed. In much of the early hype for autonomous vehicles, this idea was offered. With accidents like the one described in the Uber self-driving vehicle fatality [25], the likelihood that this assumption is valid in the near term is low.

3) Example 3 – Backup Human Monitor Interacting with Uber Autonomous Self-Driving Vehicle:
A pedestrian was killed on March 18, 2018 when an Uber-owned self-driving car operating in autonomous mode struck her as she was crossing a road in Tempe, Arizona. The self-driving car had a human safety operator who was required to pay attention to the vehicle and the environment in order to take control from the automation when the vehicle encounters difficult or unknown situations. It was later determined that the vehicle operator spent 36% of the drive watching a television show on her cell phone not paying attention for potential vehicle anomalies [25-27].

Though the 3-member team of investigators from the National Transportation Safety Board (NTSB) found six failures that contributed to the fatality, this paper would like to focus on the first failure identified:

Failure 1. The failure of the Uber self-driving vehicle operator was due to the fact that she was visually distracted throughout the trip by her personal cell phone. Had the vehicle operator been attentive, she would likely have had sufficient time to detect and react to the crossing pedestrian to avoid the fatality.

In the analysis of the fatality, NTSB also faulted Uber for their inability to address what they believed was one of the underlying reasons why the vehicle operator was not paying attention to the autonomous vehicle, namely, "automation complacency." The NTSB report stated that it was Uber's inability to address the "automation complacency" of its safety drivers who monitor the automated driving systems. They also expressed that Uber lacked a system in place to ensure its safety drivers weren't getting overly complacent due to dullness or boredom.

Automation complacency is defined as insufficient attention to monitor automation output because the output is viewed as reliable [30]. In human factors research, automation complacency is closely linked with automation bias (the tendency to trust decision-support systems). "Although the concepts of complacency and automation bias have been discussed separately as if they were independent," writes one expert, "they share several commonalities, suggesting they reflect different aspects of the same kind of automation misuse." Some have proposed that the concepts of complacency and automation bias be combined into a single "integrative concept" because these two concepts "might represent different manifestations of overlapping automation-induced phenomena" and because "automation-induced complacency and automation bias represent closely linked theoretical concepts that show considerable overlap with respect to the underlying processes" [28].

We posit that some forms of automation complacency and automation bias can be combined and coupled. We identify this coupling as a moral hazard of autonomy. Using the moral hazard framework, we differ from the NTSB perspective that the human safety operator suffered from insufficient attention to the automation. This reasoning implies that the human simply cared more about television entertainment than protecting human lives. We disagree that the human operator cared more about her own personal entertainment. We suggest an alternative explanation that the human safety operator shifted the safety risk to the automation. Why would we come to this conclusion? As the NTSB report expresses, after spending about 60% of the drive experiencing the safety and correctness of Uber's self-driving vehicle, the human operator felt comfortable shifting her risks to the automation. The operator's confidence with the autonomous technology aided her to feel insured against loss. While driving around in the Uber self-driving vehicle, the human operator rarely had to intervene, which testified to the accuracy and correctness of the technology. The more the autonomous vehicle revealed itself to be reliable in real-world driving, the more trust the human operator placed in it. This bidirectional relationship increased until the human operator felt so safe with the autonomous vehicle's capabilities that she felt free to focus her attention elsewhere because of the high confidence or trust she placed in the machine. It has been said that human beings are naturally predisposed to trust due to our genetic makeup and societal training. This willingness to trust gets us into trouble sometimes whether that trust is placed with another human or with a machine. Although humans have well-developed (but imperfect) systems for detecting untrustworthy humans, humans are wholly unprepared for the task of detecting untrustworthy machines. It is unreasonable to expect that a human will have this ability simply because they are told to do so.

The advantage of the moral hazard of autonomy explanation is that it provides a more reasonable explanation of causal events that are relatable and understandable. Additionally, the moral hazard framework ties the operator's behavior to an already established set of principles and behaviors that manifest in other fields. Despite the speculative nature of the topic, it would be interesting to investigate where moral hazard can be used to explain very perplexing and perhaps paradoxical human behavior with automation in safety-critical contexts. Because of the ironies and paradoxes that exist in human-machine teaming, some researchers believe that the human operator's perception of automation's reliability should be calibrated as a design mitigation [29]. This could also be viewed as a potential mitigation to avoid moral hazard. This calibration of the automation would be designed to maintain human workload at an appropriate level while ensuring that the human operator remains engaged with the monitoring task [29].

VI. Conclusions

This paper describes the concept of moral hazard as applied to technologies that incorporate automation and autonomy. Moral hazard is said to exist when a party to a transaction feels more comfortable taking undue risks because another party will bear the costs if things go badly. As opposed to regular physical hazards, a moral hazard comes from within a person. In this paper, we reveal two categories of moral hazards related to autonomy. The first category of moral hazard occurs when the owner of the autonomy introduces an autonomous system without accepting the full responsibility for improper operation. The issue becomes more complicated in cases where the operator of the autonomy may not have a full understanding of the system behavior. This situation is affiliated with adverse selection where there is an information imbalance between the seller and purchaser. This category of moral hazard is similar to moral hazards experienced in other industries and can often be addressed through appropriate policy and establishing liability for irresponsible behavior. In the second category of moral hazard, risks are shifted from people to automation/autonomy. In this category, the humans in proximity to the autonomous system begin to trust its behavior. Their behavior may change in that they may believe they are more insulated from harm and subsequently exhibit more risky behavior toward increasingly autonomous technologies. Mitigating this type of moral hazard may require the autonomy to possess certain design features to discourage this type of human behavior. Understanding the roles and responsibilities of the overall system and the humans is important so that humans do not suffer needlessly in their interactions with autonomous systems by placing inappropriate trust where that trust is neither warranted nor deserved.

References

[1] Perrow, C., 1984. Normal Accidents – Living with High Risk Technologies, New York, Basic Books.

[2] Murray, C., "Automakers Are Rethinking the Timetable for Fully Autonomous Cars," PlasticsToday, 17 May 2019, URL: https://www.plasticstoday.com/electronics-test/automakers-are-rethinking-timetable-fully-autonomous-cars/93993798360804 [retrieved 25 Nov. 2019].

[3] Bush, V., July 1945, "As We May Think," The Atlantic Monthly, 176 (1): 101–108.

[4] Turing, Alan (October 1950), "Computing Machinery and Intelligence," Mind, LIX (236): 433–460, doi:10.1093/mind/LIX.236.433

[5] Asimov, I., "Robbie" Part of I, Robot. Bantam Spectra 1950, Note: the Robbie story was originally published in 1940.

[6] Crevier, D., "AI: The Tumultuous History of the Search for Artificial Intelligence," BasicBooks. New York. 1993.

[7] Holmes, A., "Robots surpassed humans at these tasks in the past decade," Business Insider, 16 Nov. 2019, URL: https://www.businessinsider.com/robots-surpassed-humans-at-these-tasks-in-past-decade-2019-11 [retrieved 27 June 2020].

[8] Gvoth, W., and Goldhush, D. (2020, April 28). DABUS Denied: Only Natural Persons can be Named as Inventors on US Patents. Global IP & Technology Law, https://www.iptechblog.com/2020/04/dabus-denied-only-natural-persons-can-be-named-as-inventors-on-us-patents/#:~:text=On.

[9] Tapscott, R. (2020, May 4). USPTO Shoots Down DABUS' Bid For Inventorship. IPWatchDog, https://www.ipwatchdog.com/2020/05/04/uspto-shoots-dabus-bid-inventorship/id=121284/

[10] Moore, J., "AI, autonomous cars and moral dilemmas," TechCrunch, 19 Oct. 2016, URL: https://techcrunch.com/2016/10/19/ai-autonomous-cars-and-moral-dilemmas/ [retrieved 6 June 2018].

[11] Winston, M. (2008, January 30). Moral Agency and Autonomy. An Ethics of Global Responsibility, http://ethicsofglobalresponsibility.blogspot.com/2008/01/moral-agency-and-autonomy.html

[12] Brat, G., Davies, M., Koelling, J., Maddalon, J., and Miner, P., "Moving the Validation and Verification Frontier: the System-Wide Safety March Towards Scalability and Autonomy," AIAA Aviation 2019 Forum,

Dallas, Texas, 17-21 June 2019, DOI: https://doi.org/10.2514/6.2019-2834.

[13] National Academies of Sciences, Engineering, and Medicine 2020. Advancing Aerial Mobility: A National Blueprint. Washington, DC: The National Academies Press. https://doi.org/10.17226/25646.

[14] System-Wide Safety Project Description, NASA Headquarters, Washington, D.C., Aug. 31, 2018, URL: https://www.nasa.gov/aeroresearch/programs/aosp/sws-project-description.

[15] Morris, A.T., and Massie, M., "The Integrated Hazard Analysis Integrator," AIAA InfoTech@Aerospace Conference 2009, Seattle, Washington, 6-9 April 2009, DOI: 10.2514/6.2009-1944.

[16] Wraight, Patrick, "Is it a MORAL or MORALE hazard?" Insurance Journal, 22 Jan. 2020, URL: https://www.insurancejournal.com/blogs/academy-journal/2020/01/22/555435.htm, [retrieved 3 March 2020].

[17] "Moral hazard," Wikipedia, Wikimedia Foundation, 25 June 2020, URL: https://en.wikipedia.org/wiki/Moral_hazard.

[18] [18] Morris, A. Terry, "On the Migration of Risks and Liabilities for Increased Automation," AIAA SciTech 2020 Forum, Orlando, Florida, 6-10 January 2020, DOI: 10.2514/6.2020-0454.

[19] "Principal–Agent Problem," Wikipedia, Wikimedia Foundation, 25 June 2020, URL https://en.wikipedia.org/wiki/Principal-agent_problem.

[20] Thoma, M., "Explainer: What is "moral hazard"?," CBS News, 22 Nov. 2013, URL: https://www.cbsnews.com/news/explainer-moral-hazard/ [retrieved 25 Nov. 2019].

[21] Bernstein, P. L., "The Moral Hazard Economy," Harvard Business Review, July–August 2009 Issue, URL: https://hbr.org/2009/07/the-moral-hazard-economy.

[22] Schwalbe, M., "The Moral Hazards of Capitalism," CounterPunch, Sept. 1, 2015, URL: https://www.counterpunch.org/2015/09/01/the-moral-hazards-of-capitalism/ [retrieved 25 Nov. 2019].

[23] Adams, J., "Moral hazard and sacred cows: Will driverless cars really make us safer?," CityMetric, 5 Sept. 2016, URL: https://www.citymetric.com/horizons/moral-hazard-and-sacred-cows-will-driverless-cars-really-make-us-safer-2408 [retrieved 25 Nov. 2019].

[24] Efrati, A., "Waymo's Big Ambitions Slowed by Tech Trouble," The Information, August 28, 2018.

[25] Lee, D., "Uber self-driving crash 'mostly caused by human error'," BBC News, 20 Nov. 2019, URL: https://www.bbc.com/news/ technology-50484172 [retrieved 25 Nov. 2019].

[26] 'Inadequate Safety Culture' Contributed to Uber Automated Test Vehicle Crash, National Transportation Safety Board, 19 Nov. 2019, https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx [retrieved 25 Nov. 2019].

[27] Hawkins, A.J., "Uber is at fault for fatal self-driving crash, but it's not alone," The Verge, 29 Nov. 2019, URL:https://www.theverge.com/2019/11/19/20972584/uber-fault-self-driving-crash-ntsb-probable-cause [retrieved 25 Nov. 2019].

[28] Parasuraman, R. and Manzey, D., (June 2010). "Complacency and Bias in Human Use of Automation: An Attentional Integration". The Journal of the Human Factors and Ergonomics Society. 52 (3): 381–410. doi:10.1177/0018720810376055. PMID 21077562. Retrieved 11 July 2020.

[29] Bainbridge, L., (1983). "Ironies of automation," Automatica. 19 (6): 775–779. doi:10.1016/0005-1098(83)90046-8.

[30] Goddard, K.; Roudsari, A.; Wyatt, J. C. (2012). "Automation bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," Journal of the American Medical Informatics Association. 19 (1): 121–127. DOI:10.1136/amiajnl-2011-000089.