ON THE MORAL HAZARD OF AUTONOMY

A. Terry Morris, PhD, Jeffrey M. Maddalon and Paul Miner, PhD

Director, AIAA Information Systems Group

Sub Project Manager, System-Wide Safety Project

NASA Langley Research Center

October 13, 2020

MOTIVATION

- What do we want?

  To build increasingly autonomous (IA) systems that we design, fully understand and maintain so that they may service humanity with minimal loss of human life.

- How do we perform this function?

  We should guide the design of IA systems where we understand all the hazards that exist including appropriate mitigations against those hazards so that the autonomy services humanity appropriately.

OUTLINE

- I. AI & Autonomy Surpassing Humanity
- II. Investigating the Hazards of Autonomy
  - I. Systems Safety Hazards
  - II. Any Other Hazards?
- III. Moral Hazard
  - I. Definition of Moral Hazard
  - II. Manifestations in Other Industries
- IV. The Moral Hazard of Autonomy
  - I. Type 1 – Risk is Transferred from One Party to Another
  - II. Type 2 – Risk is Transferred from People to Autonomy
- V. Examples
- VI. Mitigations & Conclusions

AI & AUTONOMY SURPASSING HUMANITY

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, an 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
2020	Fighter Pilot	AI easily beats Human F-16 fighter pilot in DARPA Alpha Dogfight Trials

INVESTIGATING THE HAZARDS OF AUTONOMY

- System Safety involves risk-based strategies to identify, assess, eliminate and mitigate various hazards of a system which includes the importance of interactions between system components and the environment to an acceptable level of safety.
- Systems that incorporate autonomy are evaluated using tools such as functional hazard assessment, fault trees, FMECA, etc.
- Are there any ***Other Hazards*** not identified by system safety tools and techniques?

Four Ways to Manage Hazard Risks	
Avoid the Risk	Design the system and constrain its operations so that no loss occurs
Reduce (or Mitigate) the Risks	Steps are taken to identify the risks and then reduce the chance of the risks from occurring
Accept the Risks	A person can decide to assume responsibility for the risk
Transfer the Risks (2 ways)	Due to injury from negligence, transfer risks to negligent party (sue negligent party)
	Transfer risk to Insurance where the risk is shared among a number of insureds

WHAT IS MORAL HAZARD?

- 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.
- 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 specifically when humans feel they are insured against losses.

MORAL HAZARDS (IN OTHER INDUSTRIES)

Industry	Moral Hazard Exists…	Behavior when no Mitigation Exists for Moral Hazard
Automobile and Home Insurance Industries	When there are no deductible or copays in the insurance policy	Humans would likely take more risks leading to minor car damage or property theft in their homes knowing the damage is fully covered
Financial Industry	When bankers are provided government-backed insurance against their losses	"Too Big To Fail" banks can take huge risks and be bailed out thus avoiding the consequences to their bad decisions
Health Care Industry	When people only sign up for health care when they are sick	(In US) Health care system would be too costly to sustain (Note: MH mitigated very differently in countries with universal health care)

- In the Technology Industry, the ***moral hazard of autonomy*** exists when humans transfer safety risks (hazards) to other humans (type 1) or to autonomous systems (type 2). Additionally, the transfer of these risks may occur consciously (deliberate) or unconsciously (unintentional).

THE MORAL HAZARD OF AUTONOMY

- Type 1 – Risks are Transferred from one Party to Another Party

*

The owner of an autonomous system does not accept full responsibility for improper operation and does not share known product deficiencies with purchaser (risk transferred to another party).

- Type 2 – Risks are Transferred from Humans to Autonomous System

^

When autonomy functions safely and reliably over time, humans may begin to put their trust in autonomy believing they are more insulated from harm and subsequently exhibit more risky behavior toward autonomy (risk transferred from humans to autonomy).

* Image retrieved from https://encrypted-tbn0.gstatic.com/images?q=tbn%3AANd9GcThUS1vDmu1Krm0f4rwsoRBOZGCTK8CUfo1kQ&usqp=CAU

^ https://www.advancedsciencenews.com/wp-content/uploads/2019/06/hand-3308188_1920.jpg

NEW DRONE PURCHASED FOR USE IN A PARADE

- Example 1 – Billy is a high school student and participates in the marching band. He has a permit to fly unmanned drones over public airspace. Billy routinely participates in the drone challenge at his school. Billy purchases an autonomous drone from a local seller to capture footage of the parade. The seller (eager to sell the drone) does not reveal the loss of navigation problem to Billy and basically guarantees the safety of the drone. After purchasing this drone, Billy programs the drone to autonomously lift off, fly above the parade and take pictures using known GPS coordinates. Unfortunately, the drone experiences loss of navigation and lands on the head of Melissa (the flute player in the band). Melissa suffered some lacerations and was taken to the hospital.
- Analysis:
  - Type 1 Moral Hazard (Transferring Risks from One Party to Another Party)
  - Seller had prior information on deficiencies and did not share with Billy (information asymmetry)
  - Billy shifted the burden of public safety onto the seller (without questions or proof or V&V)
  - Melissa suffered the consequences of a physical collision hazard
  - Billy suffered a moral hazard that happened to incorporate autonomy

* Image retrieved from https://encrypted-tbn0.gstatic.com/images?q=tbn%3AANd9GcS4OZw1GAz_-2k-SJln8UiC1MQH5vCAltEQBw&usqp=CAU

HUMAN DRIVER INTERACTION WITH RISK AVERSE WAYMO AUTONOMOUS VEHICLE

- Example 2 – Some versions of Waymo's ride-sharing self-driving cars in Chandler, Arizona have very risk averse driving styles (perspectives from some human drivers on the road). Out of an abundance of caution, the Waymo vehicle stops abruptly in situations where no human driver would expect causing the driver following the Waymo vehicle to suddenly apply brakes. The Waymo vehicle also experiences extra-long waits for situations to clear. Some accidents where a human driver rear-ended a Waymo vehicle have been reported. One human driver illegally drove around the Waymo vehicle when it waited too long at an intersection.

*

- Analysis:
  - Type 2 Moral Hazard (Induced by the Waymo Vehicle, Human transfers risk to Autonomous Car)
  - Human driver knew Autonomous Vehicle was risk averse (therefore very safe), so human driver transferred the risk of his safety to the Waymo vehicle knowing the vehicle was designed to not crash into him (even as he was performing an illegal move).

* Imaged retrieved from https://miro.medium.com/max/579/1*BGJSKe95j7S7be8ZjTuSnA.png

HUMAN DRIVER INTERACTING WITH UBER AUTONOMOUS SELF-DRIVING VEHICLE

- Example 3 – 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 the street in Tempe, Arizona. The Uber self-driving vehicle 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 encountered a difficult or unknown situation. It was later determined that the safety operator spent 36% of the drive watching a television show on her cell phone not paying attention for potential anomalies. Among the 6 failures identified by **NTSB**, the first failure was assigned to the vehicle safety operator because she was visually distracted by her personal cell phone or suffered "automation complacency." Uber was also faulted for its inability to address "automation complacency" to ensure that its drivers don't become overly complacent due to dullness or boredom. This implies the safety driver cared more about entertainment than saving human lives.
- Analysis:
  - Type 2 Moral Hazard (Human safety operator transfers safety risk to Autonomous Vehicle)
  - Safety operator spent 60% of drive experiencing the safety and correctness of the Uber vehicle. Safety operator felt comfortable shifting the safety risks to the Autonomous Vehicle.

COMPARING CONCLUSIONS WITH NTSB

NTSB post analysis…

- The Uber safety operator suffered from automation complacency.
- Automation complacency is closely associated with automation bias.
- These two concepts are not independent, they represent closely-linked theoretical concepts that show considerable overlap.
- Their commonalities represent different manifestations of automation misuse.

Our Conclusions…

- Some forms of automation complacency and automation bias can be combined and coupled.
- We identify this coupling as the moral hazard of autonomy.
- We differ with the NTSB perspective that the safety operator cared more for entertainment than saving human lives. We posit that the Uber vehicle exhibited safety and correctness to such an extent that the safety operator placed her trust (transferred the safety risk) to the autonomous vehicle. The operator felt so safe that she focused her attention elsewhere.

POTENTIAL MITIGATIONS…

- Mitigations include 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.
- Mitigations also include monitoring, consulting experts, transacting 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.
- Mitigating some areas of moral hazard may require the autonomy to possess certain design features to discourage humans from blindly trusting its behavior.
- Further research will investigate whether particular mitigations are more appropriate when the risk is transferred consciously (deliberate) or unconsciously (unintentional).

CONCLUSIONS

- 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 we have well-developed (but imperfect) systems for detecting untrustworthiness, humans are wholly unprepared for the task of detecting untrustworthy machines.
- 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.
- When 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. We don't want humans to suffer needlessly in their interactions with autonomous systems by placing inappropriate trust where that trust is neither warranted nor deserved.