Lessons learned from self-driving cars implementation in the US

M.L. Cummings

I was recently the Senior Safety Advisor at the US National Highway Traffic Safety Administration (NOV 21 – DEC 22). As a result of this work, I offer up five lessons learned regarding self-driving cars, which are also more broadly applicable to AI in safety-critical systems.

1. Human errors in operation get replaced with human errors in coding

The AV industry routinely asserts that the sooner we get rid of drivers, the safer we will all be on roads since 94% of accidents in the US are caused by human drivers, a statistic that is taken out of context and not accurate [1]. Moreover, this claim ignores what anyone who has ever worked in software development knows all too well - coding is incredibly error-prone, especially as systems grow in complexity.

Several AV accidents with coding errors as causal factors have resulted in defect recalls including the crash of a Pony driverless car in October 2021 into a sign, a TuSimple semi-tractor trailer crash in April of 2022 into a highway jersey barrier, a Cruise crash caused by a sudden stop in the middle of an unprotected left turn in June of 2022, and another Cruise AV colliding into the back of a bus in March of 2023 [2].

These problems illustrate that we are only shifting the human error mode from the last link in the causal chain to the first few links in the chain of creating AI. This shift can be much more dangerous since such errors are latent and far harder to mitigate. Testing in simulation but predominantly in the real world is the key to reducing the probability that such errors will materialize, especially in safety-critical systems. However, without government regulation and clear industry standards, AI companies will cut corners and focus on getting products to market quickly, at the expense of reducing software errors.

2. Failure modes can be surprising

AI decisions based on neural networks is effectively probabilistic reasoning, meaning the choices of image labels and actions for a self-driving car are guesses from previous model training on pre-existing data. Failure modes are extremely hard to predict since it is impossible to precisely to know how any neural network will respond when presented with data from the real world. AVs can behave very differently on the same stretch of road at different times of the day, possibly due to sun angles, although we still don't know if this is an absolute cause.

One failure mode not previously anticipated by experts in the AV industry is the prevalence of phantom braking, where AVs execute a hard braking maneuver for an unknown reason, often leading to one or more following cars crashing into them. Such unintended hard braking maneuvers have been seen in many different manufacturer’s self-driving cars, and also in driving-assist equipped cars.

The cause of such events is still a mystery to industry and the government. This failure mode was initially attributed to humans following too closely (often followed by references to the misleading 94% statistic). However, as an increasing number of these crashes have been reported to NHTSA [3] and with the release of Tesla hard braking data in Germany [4], it is clear that the AI is not performing as it should. Moreover, this is not just one company's problem – all companies leveraging computer vision are susceptible to this problem.

As other kinds of AI begin to infiltrate other areas of society through generative AI, it is extremely important for standards bodies and regulators to be aware that AI failure modes will not follow a predictable path. They should also be wary of the propensity to excuse away bad tech behavior and incorrectly blame humans for abuse or misuse of the AI.

3. Probabilistic estimates do not approximate judgment under uncertainty

Ten years ago, there was fear of job loss over the rise of IBM's AI-based Watson, the precursor to today's large language models like ChatGPT. These fears were not realized and eventually it became evident that while Watson, as well as LLMs today, was good at making probabilistic guesses, it had no real knowledge, especially when it came to making judgments under uncertainty. This means that a decision maker has to decide on an action based on imperfect or incomplete information. LLMs and other forms of AI will not replace humans in jobs where there is decision making under significant uncertainty because the underlying models simply cannot cope with a lack of information and do not have the ability assess whether their estimates are good enough in the context at hand [5].

Indeed, these problems are routinely seen in the self-driving world. The unprotected left accident for Cruise happened when one its cars decided to make an aggressive left turn between two cars. The car correctly chose a feasible path but then halfway through the turn, slammed on its brakes and stopped in the middle of an intersection because it guessed an oncoming car in the right lane was going to turn (which was not physically possible at the speed the car was going [6]). The uncertainty in this situation effectively confused the car and it made the worst possible decision.

Cruise vehicles have also had many problematic interactions with first responders, who by default are operating in areas of significant uncertainty. These interactions have included traveling through active fire-fighting and rescue scenarios, including driving over downed power lines [7]. In one incident, a fire fighter knocked out the window of a Cruise AV to get it out of the scene. Waymo, Cruise's main rival, has also experienced similar problems [8].

These incidents show that just because an AV can access neural networks that classify a large set of images and propose a set of actions in common settings, they struggle to perform even basic operations when the world does not match the underlying training data. The same will be true for LLMs and other forms of generative AI. Just because a neural network has access to billions of pieces of information, this does not equate to judgment in the face of uncertainty, which is a key precursor to actual knowledge.

4. Maintaining AI is just as important as creating AI

Because neural networks can only be effective if they are trained on significant amounts of relevant data, their reliance on quality data is paramount. However, models cannot just be trained once and then expected to maintain high quality performance. To be useful in dynamic settings like driving, models must be constantly updated to reflect new cars, construction zones, new types of bikes and scooters, etc.

Recently a Cruise AV hit the back of an articulated bus, which was surprising since such accidents were thought to be nearly impossible for a system that carried LIDAR, RADAR, and computer vision. Cruise attributed this accident to a faulty model where the AV estimated the back of the bus based on a normal bus instead of an articulated bus, and also hampered by the rejection of the LIDAR data that correctly detected the bus.

This example highlights the importance of maintaining AI, particularly the currency of models. Model drift, a known problem in AI, occurs when relationships between input and output data change over time. For example, if a self-driving car fleet operates in one city with one kind of bus, and then the fleet moves to another city with different bus types, the underlying model of bus detection will likely drift, which could lead to very serious consequences. Maintaining model currency is just one of many ways AI requires periodic maintenance [9], and any discussion of AI regulation in the future must address this critical aspect.

5. AI should be implemented with an understanding of system-level implications

One of the most significant current problems with self-driving cars is how to deal with them when they freeze. These cars are designed to stop when they cannot resolve uncertainty (see #3), which is an important safety feature. However, recent operations for both Cruise and Waymo have demonstrated that managing such stops has been an unexpected challenge.

When these cars stop, they often block roads and intersections, sometimes for hours, including normal traffic and also first response vehicles. Companies have instituted remote monitoring centers and rapid action teams to mitigate the congestion and confusion caused by these vehicles, but the quality of their responses has been called into question by city officials in San Francisco [10]. It is also not clear what role connectivity plays in the ability to respond quickly enough, as 20 Cruise vehicles caused a massive traffic jam in part due to a loss of connectivity with the remote operations center [11].

While such growing pains are not unexpected when technology is deployed for the first time, it is also critical that companies take more care to understand the derivative problems that their technologies are likely to cause. Sentiment towards self-driving cars, which used to be optimistic, has taken a negative turn in San Francisco and the large US population. Such sentiment could be insurmountable if, for example, stopped AVs are ever implicated in the death of a person who could not get to the hospital in time.

While it is important for companies to ensure they understand broader systems-level implications, it is also equally important that regulatory agencies work to define reasonable operating boundaries for systems with AI, and accordingly, issue permits and regulations. When the use of AI presents clear safety risks, agencies should not defer to industry for solutions and should be proactive in setting limits.

These five lessons learned show that AI still has a long way to go before it can be a considered a success in vehicles. There are clear benefits to this technology if it can be introduced responsibly, but we need to have more informed conversations about such regulation with people that have technical competence in AI.

[1] H. Yen and T. Krisher, "NTSB chief to fed agency: Stop using misleading statistics," in *AP News*, 2022.

[2] NHTSA, (2021). *First Amended Standing General Order 2021-01*. US Department of Transportation.

[3] ODI, (2022). *PE 22-002, Unexpected activation of braking system may cause rapid deceleration*. US Department of Transportation.

[4] Staff, ""My autopilot almost killed me": Tesla files cast doubt on Elon Musk's promises," *Handelsblatt*. [Online]. 2023. Available: https://www.handelsblatt.com/unternehmen/industrie/elektromobilitaet-mein-autopilot-hat-mich-fast-umgebracht-tesla-files-naehren-zweifel-an-elon-musks-versprechen/29166564.html

[5] M. L. Cummings, "Rethinking the maturity of artificial intelligence in safety-critical settings," *Artificial Intelligence Magazine,* vol. 42, no. 1, pp. 6-15, 2021.

[6] M. Woon, "Unacceptably Risky – Part 1: Cruise's Own Allegations Following the June 3rd Crash Establish a Reasonable Doubt," Retrospect Consulting, 23 November 2022.

[7] S. Whiting, "Self-driving Cruise cars that tangled with S.F. Muni lines had no passengers, company says," *San Francisco Chronicle*. [Online]. 2023. Available: https://www.sfchronicle.com/sf/article/cruise-muni-sf-storms-driverless-17856051.php

[8] J. Eskenazi, " 'No! You stay!' Cops, firefighters bewildered as driverless cars behave badly," *Mission Local*. [Online]. 2023. Available: https://missionlocal.org/2023/05/waymo-cruise-fire-department-police-san-francisco/

[9] M. L. Cummings, "Revising human-systems engineering principles for embedded AI applications," *Frontiers in Neuroergonomics: Social Neuroergonomics,* vol. 4, 2023, doi: 10.3389/fnrgo.2023.1102165.

[10] T. Claburn, "City isn't keen on 5,000 erratic, traffic-jam-causing GM robo-cars on its streets," *The Register*. [Online]. 2022. Available: https://www.theregister.com/2022/09/27/gm_cruise_robocar_safety_waiver/

[11] A. Marshall, "Cruise's Robot Car Outages Are Jamming Up San Francisco," *Wired*. [Online]. 2022. Available: https://www.wired.com/story/cruises-robot-car-outages/