Shared Authority Concerns in Automated Driving Applications

Prof. M.L. Cummings & Jason Ryan
Massachusetts Institute of Technology
Humans and Automation Laboratory

Given the move toward driverless cars, which includes the more short-term goal of driving assistance, what the appropriate shared authority and interaction paradigms should be between human drivers and the automation remains an open question until more principled research and testing has occurred. It is unclear at this time how robust driverless cars are to system failures (including human failures) and operations in degraded sensor environments. Automation onboard such vehicles is inherently brittle and can only account for what it is programmed to consider. Communication between what is technically a very complex system to a human population of extreme variability in driving skills and attention management will be difficult, since the driver will need to be appropriately informed of the state of the system, including limitations, and will need to build appropriate trust in the automation's capabilities (neither too much or too little). Further complicating this problem is the significant body of research demonstrating that automated systems can lead to boredom, which encourages distraction. This leaves operators unaware of the state of the vehicle (aka, mode confusion) and ill-suited to respond quickly and appropriately in case of a potential accident. Over time, operator skill degradation due to automation use can further reduce the human ability to respond to emergent driving demands, and will likely lead to risk homeostasis even in normal operations. Each of these issues are well-known to the human systems engineering community, but it is unclear that these issues are being considered by driverless car designers or that manufactures are conducting human-in-the-loop tests with representative members of the driving population. Until these tests show that the vehicles account for the aforementioned issues, driverless cars will not be safe for unrestricted access and use on U.S. roadways. Moreover, there are significant socio-technical considerations that do not appear to be a concern in the push to introduce this technology on a wide scale. The utilitarian approach quoted by many in the press, i.e., that driverless cars will eventually kill people but that this should be acceptable due to the likely reduction in overall deaths (which is not yet proven) demonstrates an insensitivity to a deontological perspective that causes many people to be uncomfortable with such a significant shift in responsibility and accountability to computers.

Driverless is Really Driver-Optional

Google, Volvo, GM, Audi, Toyota and other companies designing and testing driverless" cars claim that they are capable of navigating roadways, changing lanes, observing traffic signals, and avoiding pedestrians without human input. They do so through a combination of technologies including GPS position information, internal navigation maps, outward-facing cameras, and the use of laser (and other) range-finding systems. The first two of these technologies allow the vehicle to understand where it is in the world where it should be going, and how to get there; the latter two allow the vehicle to track where it is on the road and where other vehicles, traffic indicators, and pedestrians are.

While termed "driverless," the vehicles are better classified as driver-optional, particularly under NHTSA Levels 2 and 3 of automated driving, where human operators are expected to have either primary or secondary control responsibilities. Although such vehicles are supposedly capable of driving through any traffic situation without requiring a human driver to apply pressure to the pedals, shifting, or steering, this driver may still choose to do so and may play a role in avoiding accidents. Google readily admits that, while testing, there remains a safety driver and a software operator in the vehicle at all times in the case of near accidents or software failures (Thrun 2010).

While ultimately in the distant future, this driver won't be needed, these systems currently require a human to be in the driver's seat and allow (and in some cases, expect) the driver to assume control at specific points in time. It is here that the problem lies: as long as a human operator has some expectation of shared authority, either primary or secondary, the design of the automation must be such that the operator fully understands the capabilities and limitations of the vehicle and maintains full awareness of what the system is doing. Failure to do so may lead to a variety of accidents, both automation and human-induced.

While Google's driverless cars have already logged over 300,000 miles (Urmson 2012), two accidents have occurred. One of these occurred while the car was traveling on roads not previously mapped into its system (DeBolt 2011). This is evidence of automation brittleness in that the car was not able to handle the uncertainty in its internal model, exacerbated by human error (CBS News 2011). Such problems are further exacerbated by an inherent human limitation known as neuromuscular lag (Jagacinski and Flach 2003), where even if a person is paying attention perfectly, there is still an approximate half second lag for a person to see a developing situation and then take action accordingly. Instances of "human error" like this are not the fault of the human alone; it is instead a fault of the interaction between the human and the automation and their respective weakness of imperfect attention and response execution (for the human) and brittleness in perception and solution generation (automation.)

Lessons from the Aviation Community

The driverless car community can look to aviation for many lessons learned from the introduction of automation to relieve pilot workload and in theory, improve safety. Many accidents labeled as human error by the FAA and the NTSB are better categorized as failures of human-automation interaction (Dismukes, Berman et al. 2007).

- The 1972 crash of Eastern Airlines flight 401 was caused by a faulty oil warning light that appeared on final approach. The crew, distracted by the disagreement between the warning light and other gauges, failed to notice that the autopilot disengaged, with no warning to the pilots. They slowly descended into the Everglades.
- Air France 447, which crashed off the coast of Brazil in 2009, involved two failures: Failure of the automation, and a failure to present information to the operator. A clogged pressure sensor lead the autopilot system to sense the airplane was too low in altitude. The autopilot then put the aircraft into an increasing high climb, eventually triggering the stall warning alert. With the aircraft on autopilot, the pilot allowed himself to become distracted. When the stall warning activated, the pilot was not aware of what was happening and made the worst of all possible decisions – he attempted to further increase the aircraft's climb angle, worsening the stall, and contributing to the crash.
- Northwest Flight 188, which overshot Minneapolis, MN by roughly an hour in the fall of 2009, was caused by operator boredom and resultant distraction. With the aircraft autopilot in control, both pilots became distracted by their conversation and failed to monitor the aircraft and its status. As they opened their laptops to obtain information to supplement their conversation, they misdialed a radio frequency change, missed at least one text message sent by air traffic control to find out where they were, and only realized what was occurring when queried by a flight attendant on the landing time. Luckily, this only resulted in a late landing, but more severe consequences could easily have occurred.

These issues are common to many other domains involving human interaction with automation systems and are well known to the human factors and experimental psychology communities. In general, the research community agrees that human attention is a limited resource to be allocated, and that the human brain requires some level of stimulus to keep its attention and performance high. Lacking this input, they seek it elsewhere, leaving them susceptible to distraction either by external stimuli or by the wrong information. This means that operators may miss important cues from the automation or from the environment (Eastern Flight 401), or may see the cues but may not have all the appropriate information required to make a correct decision (Air France 447), or use their spare capacity to engage in distracting activities leading to a loss in situational awareness (Northwest 188). They might also enter a state of "mode confusion," where the operator makes decisions believing that the system is in a different state than it currently is (Lankenau 2002).

While the earlier examples and research come from the aviation domain, the role of a pilot monitoring an aircraft autopilot system differs little from the "driver" of a driverless car. Recently, research in human-automation interaction has expanded to automated driving systems and are showing the same effects (Rudin-Brown and Parker 2004, Saxby, Matthews et al. 2007, Young and Stanton 2007, Vollrath, Schleicher et al. 2011, Neubauer, Matthews et al. 2012, Jamson, Merat et al. 2013). Drivers that were placed in an autonomous or highly automated car were less attentive to the car while the automation was active, more prone to distractions (especially using their cellular phone), slower to recognize critical issues, and slower to react to emergency situations (such as emergency braking). There were benefits from using automated systems with lower average speeds and better separation between vehicles during their tests, but these came at the cost of poorer performance in emergency situations. In other words, at precisely the time when the automation needs assistance, the operator could not provide it and may actually have made the situation worse. We cannot assume the operator to be always engaged, always informed, and always ready to intervene and make correct decisions when required by the automation or the situation do so. This goes for highly-trained pilots of commercial airliners as well as the general driving population of the United States and other countries (who receive little to no formal training and assessment).

Technology Robustness

Because much of the development of driverless cars is proprietary, we do not know their exact capabilities at this time. As such, we cannot make definitive statements about a specific vehicle and can only comment on the limitations of the technology overall and outline specific questions of concern. As best we understand, Google's autonomous car relies on four major technologies in its autonomous operations: LIDAR (Light Detection and Ranging), a set of onboard cameras, GPS, and stored maps in the vehicle's onboard computer. The GPS signal tells the car where it is on the stored map and where its final destination is, and from this, the car determines its route. Cameras and LIDAR help the vehicle sense where it is on the road, where other vehicles are, and where to find and follow stop signs and streetlights.

Each of these systems is vulnerable in some form or fashion and it is not clear whether any redundancy exists, or if any one of the four systems fails (maps, GPS, cameras, LIDAR), the car will not be able to function correctly. If the GPS or maps fail, the car does not know where it is on its route and where it should be going. If the LIDAR fails, it may not be able to detect other nearby cars, pedestrians, etc. If the cameras fail, it may not be able to recognize a stop sign or the current color of the traffic light. In addition, it is not clear how much advanced mapping is required by driverless cars and the frequency of map updates that are required to maintain an effective 3D world model by which the onboard computer makes decisions.

The security of GPS has also been questioned (Volpe National Transportation Systems Center 2001, Humphreys, Ledvina et al. 2008). GPS "spoofing" (mimicking the GPS signal to provide false location information) and jamming (forcibly denying GPS signal) attacks have already been observed in US military operations (Franceschi-Bicchiera 2012, Waterman 2012) as well as in civilian applications (Marks 2012). It would not be far-fetched to imagine an individual or group of individuals spoofing GPS signals in major metropolitan areas during rush hour and forcing cars off the road into buildings, off bridges, or otherwise causing damage.

Google's own researchers admit that inclement weather and construction areas are something they have yet to master (Urmson 2012). Precipitation, fog, and dust are known problems for LIDAR sensors, which can interfere with the image detection capabilities of the camera and can scatter or block the laser beams sent out by the LIDAR. Cameras are also sensitive to such problems. This leaves the vehicle unable to sense the distance to other cars and unable to recognize stop signs, traffic lights, and pedestrians. Urmson also notes that the technology cannot currently handle construction signs, traffic cops (which requires sophisticated gesture recognition that is still an immature technology), and other non-normal driving conditions. A related question is how well the system can anticipate the actions of other drivers; it is one thing to be able to avoid a car calmly changing lanes, and another entirely to anticipate the actions of a reckless and irrational driver. Given that prior research has shown that people are prone to distraction, any failures or degradations in the technology will significantly increase the likelihood of a serious or fatal accident.

Effects over Time: Trust and Skill Degradation

How drivers adapt to the presence and performance of the automation over time is not a trivial issue. If the automation is perceived to be unreliable or not proficient, then the operator refuses to use the system, regardless of any potential benefits (Parasuraman and Riley 1997). However, when automation is perceived to be proficient, operators rely more heavily on the technology and fail to utilize their own skills (Parasuraman, Sheridan et al. 2000). This leads to a loss of skill and further increases reliance on the automation (Lee and Moray 1994), possibly leading back to issues with mode confusion as previously discussed. Skill degradation due to over reliance on automation is such a problem in aviation that the FAA recently released a safety notice recommending that pilots fly more in manual mode than using the autopilot (Federal Aviation Administration 2013). Another possible concern with increasing automation is the concept of risk homeostasis (WIlde 1998) where drivers could begin to accept more risk as they perceive the automation to be more capable, which could lead to increased distraction and reliance on the automated system.

The effective prevention of these issues can be accomplished by providing appropriate feedback to the operator on their performance and the performance of the automation. This is often referred to as designing the system for “appropriate” trust (Lee and See 2004). The automation should be capable of describing its performance and its limitations to the driver, who should then be able to learn how best to use the automation in the course of their driving routine. The automation should also be able to sense when the human operator is performing poorly, or even dangerously, so that it can either support the driver or take over control. The end result is more of a partnership – each side understanding and accounting for the abilities and limitations of the other.

Socio-technical Considerations

A common argument in favor of inserting driverless car technology as soon as possible is that accidents and fatalities will be dramatically reduced, as expressed by Google’s Sebastian Thrun, “…more than 1.2 million lives are lost every year in road traffic accidents. We believe our technology has the potential to cut that number, perhaps by as much as half (Thrun 2010).” While certainly a logical argument in keeping with rational decision-making theory, such a utilitarian approach is not universally shared. A deontological approach could assert that machines should not be allowed to take the lives of humans under any circumstances, which is similar to the three laws of Asimov.

Even if the fatality rate is lower than that of human-operated vehicles (which is not a guarantee for autonomous cars, particularly for those at NHTSA levels 2/3), the idea of a machine killing a human, even accidentally, will likely not resonate with the general public. Indeed, there has been recent intense media and public campaigns against autonomous weaponized military robots (Human Rights Watch 2012). These issues will likely also be raised as significant concerns once driverless (and especially driver assisted) technology is either responsible for a fatality or a serious accident that receives intense media attention. Furthermore, the chain of legal responsibility for driverless or driver assistance technologies is not clear as well as what basic form of licensure should be required for operation. Manufacturers and regulatory agencies of driverless technologies bear the responsibility of not only considering the technological ramifications of this technology, but also the socio-technical aspects, which at this point, has not been satisfactorily addressed.

Summary

Driverless car technology is promising in terms of creating safer and more efficient driving systems, but many questions remain. As discussed, the robustness of the technology and the interaction between the human driver and driverless technology are unclear. Boredom and distraction, mode confusion, recovery from automation errors, skill degradation, and trust issues are all of major concern and have been observed in both experimental and real-life settings. However, there are solutions to these problems through proper design, supplemented by extensive testing to confirm that the solutions had the intended effect.

At this time, manufacturers have not provided, to our knowledge, any documentation describing how they have addressed these issues in their designs, including extensive and independent principled testing. Adding to the problem is the fact that these issues lie outside the typical tests performed by NHTSA in assessing safety. Until such time as these issues have been addressed through independent human-in-the-loop testing with representative user populations, these vehicles should remain experimental. We encourage NHTSA to develop such a program in their role as a safety monitor of U.S. automobiles, as well as developing a program to test the reliability and robustness of the technologies (GPS, LIDAR, etc.) and requirements for driver training, continuing education and licensure associated with these vehicles.

The development of driverless car technologies is critical for the advancement of the transportation industry. However, the majority of promises and benefits of the driverless car will likely only be realized once all cars have such advanced technologies, and we achieve NHTSA's Level 4 of fully autonomous driving. At this point in time, we are in a very tenuous period where we are attempting to transition new and unproven technologies into a complex sociotechnical system with significant variation in human ability. In addition, public perception is fast becoming a major obstacle but is surmountable. To this end, great care should be taken in the experimentation with and implementation of driverless technology as an ill-timed serious accident could have unanticipated public backlash, which could affect other robotic industries as well.

References

CBS News. (2011). "Google self-driving car crash caused by human error - says Google." Retrieved May 11, 2013.

DeBolt, D. (2011). "Google's self-driving car in five-car crash " Retrieved May 11, 2013.

Dismukes, R. K., B. A. Berman and L. D. Loukopoulos (2007). The Limits of Expertise: Rethinking Pilot Error and the Causes of Airline Accidents, Ashgate Publishing, Ltd.

Federal Aviation Administration (2013). Safety Alert for Operators 13002. F. S. Service. Washington DC, Department of Transportation.

Franceschi-Bicchiera, L. (2012). "Drone Hijacking? That's Just the Start of GPS Troubles." Retrieved April 27, 2013.

Human Rights Watch. (2012). "Arms: New Campaign to Stop Killer Robots." Retrieved May 11, 2013.

Humphreys, T. E., B. M. Ledvina, M. L. Psiaki, B. W. O. Hanlon and P. M. Kintner (2008). Assessing the Spoofing Threat : Development of a Portable GPS Civilian Spoofer. ION GNSS. Savannah, GA.

Jagacinski, R. J. and J. M. Flach (2003). Control Theory for Humans: Quantitative Approaches to Modeling Performance. New Jersey, Lawrence Erlbaum Associates, Publishers.

Jamson, A. H., N. Merat, O. M. J. Carsten and F. C. H. Lai (2013). "Behavioural changes in drivers experiencing highly-automated vehicle control in varying traffic conditions." Transportation Research Part C: Emerging Technologies **30**: 116–125.

Lankenau, J. B. a. A. (2002). A Rigorous View of Mode Confusion. SafeComp, Bremen, Germany, Springer Verlag.

Lee, J. and N. Moray (1994). "Trust, self confidence, and operators' adaptation to automation." International Journal of Human-Computer Studies **40**: 153-184.

Lee, J. D. and K. A. See (2004). "Trust in technology: Designing for Appropriate Reliance." Human Factors **46**(1): 50-80.

Marks, P. (2012). "GPS jamming: a clear and present reality." Retrieved 2013, April 27.

Neubauer, C., G. Matthews and D. Saxby. (2012). The Effects of Cell Phone Use and Automation on Driver Performance and Subjective State in Simulated Driving. Human Factors and Ergonomics Society Annual Meeting.

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

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

Rudin-Brown, C. M. and H. A. Parker (2004). "Behavioural adaptation to adaptive cruise control (ACC): implications for preventive strategies." Transportation Research Part F: Traffic Psychology and Behaviour **7**(2): 59–76.

Saxby, D. J., G. Matthews, E. M. Hitchcock and J. S. Warm. (2007). Development of Active and Passive Fatigue Manipulations Using a Driving Simulator. Human Factors and Ergonomics Society Annual Meeting, Baltimore, MD.

Thrun, S. (2010). "What we're driving at." Retrieved May 11, 2013.

Urmson, C. (2012). "The self-driving car logs more miles on new wheels." Retrieved May 11, 2013.

Vollrath, M., S. Schleicher and C. Gelau (2011). "The influence of cruise control and adaptive cruise control on driving behaviour--a driving simulator study." Accident Analysis & Prevention **43**(3): 1134–1139.

Volpe National Transportation Systems Center (2001). Vulnerability Assessment of the Transportation Infrastructure Relying on the Global Positioning System Department of Transportation.

Waterman, S. (2012). "North Korean jamming of GPS shows system's weakness." Retrieved April 27, 2013.

WIlde, G. J. S. (1998). "Risk homeostasis theory: an overview." Injury Prevention **4**: 89–91.

Young, M. S. and N. a. Stanton (2007). "Back to the future: brake reaction times for manual and automated vehicles." Ergonomics **50**(1): 46–58.