Adaptation of Human Licensing Examinations to the Certification of Autonomous Systems

M.L. Cummings
Duke University

Abstract

In aviation and surface transportation settings, pilots and drivers are certified to operate vehicles through a licensing process that includes an assessment of physical readiness, a written knowledge exam, and a practical exam, sometimes referred to as a "checkride". This is a process whereby a human examiner assesses whether humans can effectively operate autonomously either in actual or simulated conditions. Given the rise of autonomous transportation technologies and the debate as to how such technologies that leverage probabilistic reasoning could and should be certified, could such licensing approaches be extended to certification of autonomous systems? This chapter will discuss how machine and human autonomy are similar and different, how and why licensing processes have developed historically across the two transportation domains, and how autonomous transportation systems could adapt the principles of human-based licensing processes for certification.

Introduction

Recent advances in sensor development have enabled new artificial intelligence techniques in the post-processing of sensor data such as deep learning, which have greatly enhanced the perceptual abilities of autonomous systems, most notably, autonomous cars and unmanned aerial vehicles (aka drones). Both systems are expected to reach operational capabilities in the near-term, with the promise of flying cars in the distant, but realizable, future (Nneji, Stimpson, Cummings, & Goodrich, 2017).

With the imminent arrival of such vehicles, there is significant discussion around how such systems should be certified as safe to operate, given that humans will no longer be directly responsible for their operation. In aviation and surface (i.e., driving) transportation domains, operators are typically certified to control vehicles through an examination process which typically includes a physical readiness assessment, a written knowledge exam and a practical examination in the actual vehicle.

These practical exams are operational in nature and in aviation, take the form of a "checkride" whereby a human examiner assesses whether humans can effectively and safely operate an aircraft autonomously. While driving practical exams are not as focused on emergencies as aviation checkrides, they also are attempting to ensure that a vehicle will be safely operated by the human behind the wheel. In these domains, when an operator is licensed, they are effectively certified by a regulatory agency as safe to operate their vehicles across a variety of circumstances. Additional specialized licenses are needed in both driving and aviation when a person wishes to conduct commercial operations, for example, or operate vehicles under special cases.

The intensity of training and difficulty of the examination process varies with the increasing complexity of operating a vehicle. For example, it is much more difficult to gain a pilot's license for a large commercial passenger aircraft as opposed to a small four seat recreational plane. The same is true for large tractor trailers versus a standard automobile, with increasing requirements for increasing gross vehicle weights. However, the licensing of both aviation and surface transportation vehicles, commercial or otherwise, share the same basic structure in that there are three basic elements of the licensing process: 1) Basic physical standards, 2) A written exam that assesses knowledge, and 3) the practical exam where an examiner accompanies the licensee in real world conditions.

The licensing processes for aviation and surface transportation will be detailed in the following sections, followed by discussions of how such licensing processes could or should inform certification of autonomous vehicles in their respective settings. This chapter focuses primarily on autonomous cars with no required human supervision, i.e., cars with Levels 4 or 5 of automation as defined by SAE standard J3016. For clarity purposes, the term licensing will be reserved for the independent verification that a single human can effectively and safely operate a class of vehicles, while certification will mean that one or a group of homogenous vehicles (in terms of capabilities, sensors, and hardware) can be effectively and safely operated by one or more onboard computers.

Driving Licensing Exams

In the Unites States, the driver's licensing process typically begins with a vision screening test, since vision is the sense responsible for 95% of driving-related inputs (Shinar D & Schieber F., 1991). While each state has its own set of standards, typical requirements include corrected visual acuity of 20/40, but people who exceed this could be allowed a license under limited circumstances and with special on-road assessments (American Medical Association, 2003). It is not uncommon for physicians to recommend limited driving, particularly at night and in low-light conditions for drivers with diseases and conditions that affect their vision. While there are many other medical conditions that can limit a person's driving ability such as hearing loss, epilepsy and insulin use (49 CFR 391.41), waivers are given on a case by case basis. While general physicals are not required for basic licensing, they are for commercial licenses.

Once basic visual acuity is established, drivers typically take a knowledge examination, which again varies by state, but generally focuses on traffic laws, rules of the road, and safe driving practices. Once the knowledge exam has been passed, the final portion is the on-road practical exam. While each state is responsible for their own licensing programs and the exact details can vary for both personal and commercial licensing, there are common guidelines that U.S. states can reference in the design of these exams (American Association of Motor Vehicle Administrators, 2014).

These guidelines maintain that the purpose of the knowledge exam is to ensure that driver license applicants possess the information required to operate vehicles in a way that is consistent with public safety and mobility. According to these guidelines, the primary source of this information should be made readily available through a driving manual, with alternate forms of presentation for people who cannot read and other disabilities. The categories that make up the core body of testable knowledge for written driving exams is detailed in Table 1, with example topics.

The examples in Table 1 were selected to highlight areas where either humans or autonomous cars experience difficulties. For example, human drivers can struggle with anticipating stops under the vehicle control category, or performing quick stops in emergency situations. These are actions that autonomous cars perform with a much higher degree of reliability. However, autonomous cars struggle in poor lighting conditions such as sunset since sensor washout is a problem and at present, they also have difficulties operating in and around work zones due to wide variability in signage and placement of maintenance equipment. These issues will be further explored in a later section.

Table 1: Written Driving Exam Knowledge Categories with examples (American Association of Motor Vehicle Administrators, 2014)

Pre/Post driving	Trip Planning	Positioning Vehicle	Crossing/entering with vision obstructed
	Donning seat belts		Avoiding blind spot
Vehicle Control	Anticipating stops	Handling Emergencies	Quick stop
	Adjusting speed for turn		Escape paths
Rules of the Road	Shared left-turn lanes	Sharing the Road	Bicyclists
	Work zone signs		Work zones
Visual Search	Avoiding distraction	Special Driving Situations	Rural road driving
	Lane changing		Mountain driving
Communication	Uses hand signals when appropriate	Driver Preparation	Sunlight/sunset
	Horn		What to do when a driver is aggressive
Adjusting Speed	Hydroplaning	Vehicle Readiness	Displays (legibility)
	Roadside activity		Controls (ease of reach, operation)

Written knowledge exams typically assess licensees though multiple choice questions that are selected from a larger database of questions that cover the knowledge domains in Table 1. Some states administer their knowledge exams in two parts that address both general questions about the topics in Table 1, and then another section that addresses an examinee's understanding of road signs. Figure 1 is an example of road signs that examinees would have to identify.

The last element of a typical driver's licensing exam is the practical test which assesses whether an examinee has the skills to operate an automobile "in a manner consistent with the safety and mobility of the motoring public (American Association of Motor Vehicle Administrators, 2014)." The AAMVA breaks down the required skills into three categories: perceptual, attentional, and motor, which are detailed in Table 2. Per these licensing guidelines, navigation is a cognitive attribute and should not be part of skills testing, nor should advanced

Figure 1: Example signs to be identified in a written driving exam(Virginia Department of Motor Vehicles, 2017)

skills like controlled skids be assessed in any practical exam.

Table 2: Skills that should be assessed during a practical driving test (American Association of Motor Vehicle Administrators, 2014)

Attentional	**Routine Motor Skills**
Attention-sharing	Acceleration, slowing
Attention-shifting	Turning
Perceptual	Maintaining speed
Spatial judgment	Shifting
Gap judgment	Lane keeping
Distance judgment	Stopping, backing
Hazard detection	Adjusting to limited traction

These practical skills can be assessed through both on and off road tests, as well as simulation, although the on-road test with an accompanying human examiner is the most common form of driving skill assessment. Such a "check ride" where the examiner accompanies a driver through a course on public roads is an inherently subjective and uncontrolled process, where conditions will naturally vary due to weather, traffic, and the examiner's particular approach to testing. Thus it is not meant to be a scientific process that can predict who will likely crash in the future, but rather exists to verify an examinee possesses the minimal knowledge and skills necessary to drive (American Association of Motor Vehicle Administrators, 2014).

Aviation Licensing Exams

Just as driving licenses can fall into two categories, personal and commercial aviation licensing follows a similar structure. Human pilots are "rated" in various aircraft which can range from unmanned aircraft (where pilots can have minimal aviation experience) to a commercial pilot, multi-engine airline transport rating which is the most difficult to achieve and requires substantial experience. Regardless of the type of rating, pilots must also undergo physical, knowledge, and practical exams similar to that of drivers.

Physical exam requirements generally become more stringent as the size and complexity of the aircraft increases. Glider pilots, for example, only have to write a statement that they have no physical defects that would prevent them from safely operating an unpowered aircraft, but private pilots must pass a physical examination administered by an FAA-authorized aviation medical doctor before they can fly alone. This must be renewed every 5 years until pilots are 40 years of age, then they must renew every 2 years. Passenger airline transport pilots must obtain a first class medical certificate, which is a more comprehensive exam, and it must be renewed every 6 months after a pilot's 40th birthday (14 CFR Part 61).

Just as in drivers' licensing, the physical exam exists to ensure that humans are able to use their required sensors to obtain information about the world and make action decisions in a timely manner. As the FAA's requirements for increasing frequency of physicals indicates, there is enough evidence to suggest that natural aging processes can degrade a person's ability to respond both correctly and in a timely fashion in an aircraft. Because of the possible high consequences of mistakes, commercial airline transport pilots must retire at 65 years, although there is no age limit for private pilots.

Similar to driving, pilots must pass knowledge tests that demonstrate that they understand the knowledge, skill, or risk management elements defined in the Airman Certification Standards (ACS) document for the rating they are trying to achieve. The ACS also serves as the document that guides examiners during the "check ride", i.e., the practical exam. The ACS is organized such that areas for operation/tasks are listed, accompanied with information that elucidates what pilots should "Know/Consider/Do" in these circumstances (Table 3).

Table 3: Elements of the Airman Certification Standards

Expected Pilot Behaviors	Testable Elements
Know	Aeronautical knowledge
Consider	Aeronautical decision-making & special emphasis
Do	Flight proficiency

Table 4 is an ACS for a pilot undertaking the task of diverting an aircraft when the original destination is unavailable due to weather or some other unforeseen problem. It tells pilots what they need to know, what skills are required, and what considerations are needed for risk management when they are attempting to conduct a successful divert to a new destination. It also clearly underscores what errors of omission lead to failed risk management. In the ACS, a commercial pilot is responsible for 11 areas of operation that include, for example, preflight procedures, navigation, and emergency operation. In total they must exhibit knowledge, skills, and risk management abilities for at least 60 tasks.

Table 4: Airman Certification Standards for the Task of Diversion in the Area of Operation of Navigation (Flight Standards Service, 2017)

Task	***C. Diversion***
References	FAA-H-8083-2, FAA-H-8083-25; AIM; Navigation Charts
Objective	To determine that the applicant exhibits satisfactory knowledge, risk management, and skills associated with diversion.
Knowledge	The applicant demonstrates understanding of:
CA.VI.C.K1	Selecting an alternate destination.
CA.VI.C.K2	Situations that require deviations from flight plan and/or ATC instructions.
Risk Management	The applicant demonstrates the ability to identify, assess and mitigate risks, encompassing:
CA.VI.C.R1	Collision hazards, to include aircraft, terrain, obstacles, and wires.
CA.VI.C.R2	Distractions, loss of situational awareness, and/or improper task management.
CA.VI.C.R3	Failure to make a timely decision to divert.
CA.VI.C.R4	Failure to select an appropriate airport.
CA.VI.C.R5	Failure to utilize all available resources (e.g., automation, ATC, and flight deck planning aids).
Skills	The applicant demonstrates the ability to:
CA.VI.C.S1	Select a suitable airport and route for diversion.
CA.VI.C.S2	Make a reasonable estimate of heading, groundspeed, arrival time, and fuel consumption to the divert airport.
CA.VI.C.S3	Maintain the appropriate altitude, ±100 feet and heading, ±10°.
CA.VI.C.S4	Update/interpret weather in flight.
CA.VI.C.S5	Explain and use flight deck displays of digital weather and aeronautical information, as applicable.

While there is a pilot written knowledge exam similar to that in the driving domain, there is also a knowledge portion of the practical exam that takes the form of an oral exam prior to the in-flight skills check ride. The written exam tests applicants' knowledge of the rating they are applying for, and the oral exam allows evaluators to further assess missed questions on the written test through scenario-based questioning. This scenario-based questioning is especially relevant to assessment of risk mitigation since there are no set of absolute procedures for every possible contingency and emergency. This more loosely structured assessment of risk mitigation does not exist in the driving domain, and as will be discussed later, will be a key issue in the certification of autonomous systems.

Once applicants pass the written and oral exam, they proceed to the in-flight practical test which tests applicants' required skills in the performance of various tasks and maneuvers. During this portion of the exam, the evaluator can further assess not just the skills but the abilities of pilots to conduct safe risk mitigation through simulation of emergencies (such as retarding one throttle of a two engine plane to simulate an engine flameout.) The evaluator can then witness first hand as the applicant maneuvers the actual plane as well as communicate with air traffic control to effect a safe landing. The ability of a human evaluator to alter his or her test strategy in flight to adjust to differences in human performance is a key element of this evaluation process. Just as in driving, the point of these exams is not to determine predictors for eventual problems, but to ensure applicants can safely operate aircraft.

SRKE Taxonomy

The previous discussions of driving and pilot licensing procedures demonstrates that there is a clear commonality between these programs. Generally, the licensing of humans to operate vehicles focuses on the ability of the human to perceive the world around them correctly (tested through physical exams), understand the rules and guiding principles of the overarching systems (knowledge exam), and then demonstration of the skills required to safely operate in these setting (the practical exam/check ride). The one remarkable difference between driving and pilot licensing is that pilots of all levels must explicitly demonstrate risk management in their operations, where drivers do not. As will be illustrated, this should be a significant consideration in the certification and implementation of autonomous cars.

To better understand why this seemingly small difference is so important, and why autonomous vehicle (AV) certifications will need to look much more like aviation certifications than those of driving, the SRKE (skills, rules, knowledge, and expertise) taxonomy is introduced in Figure 2 (Cummings, 2014). This framework demonstrates how and when functions should be allocated between humans and autonomous systems, and what the implications are for certification of autonomous systems of all types.

In the SRKE taxonomy, skill-based behaviors are sensory-motor actions that are highly automatic, and for humans are typically acquired after some period of training (Rasmussen, 1983). In flying and driving, the bulk of a human operator's work is a set of motor responses that should become routine and nearly effortless with practice. In Figure 2, an example of skill-based control the act of flying an aircraft. Student pilots spend the bulk of their training learning to scan dials and gauges so that they can instantly recognize the state of an aircraft and adjust as needed. However, automation is superior in this task because such tasks require motor memory with a feedback error correction loop. In most commercial aircraft, automation does the bulk of flying and pilots only typically manipulate controls 3-7 minutes for flights of any length (Cummings, Stimpson, & Clamann, 2016).

There is no question that the bulk of car and aviation accidents are due to human error, often because the correct skills were not applied in a timely fashion (National Center for Statistics and Analysis, 2017; Wiegmann & Shappell, 2001). Because the automaticity of human skill-based control can be brittle due to problems like distraction, vigilance, fatigue, and the neuromuscular lag (Jagacinski & Flach, 2003), they arguably should be replaced with automation in environments requiring these skills. However, as will be discussed further, whether skill-based reasoning can be executed reliably by AVs across the breadth of required scenarios is still an open question.

Once a set of skills is mastered, both computers and operators can then attend to higher cognitive tasks called rule-based behaviors (Figure 2), which are effectively those actions guided

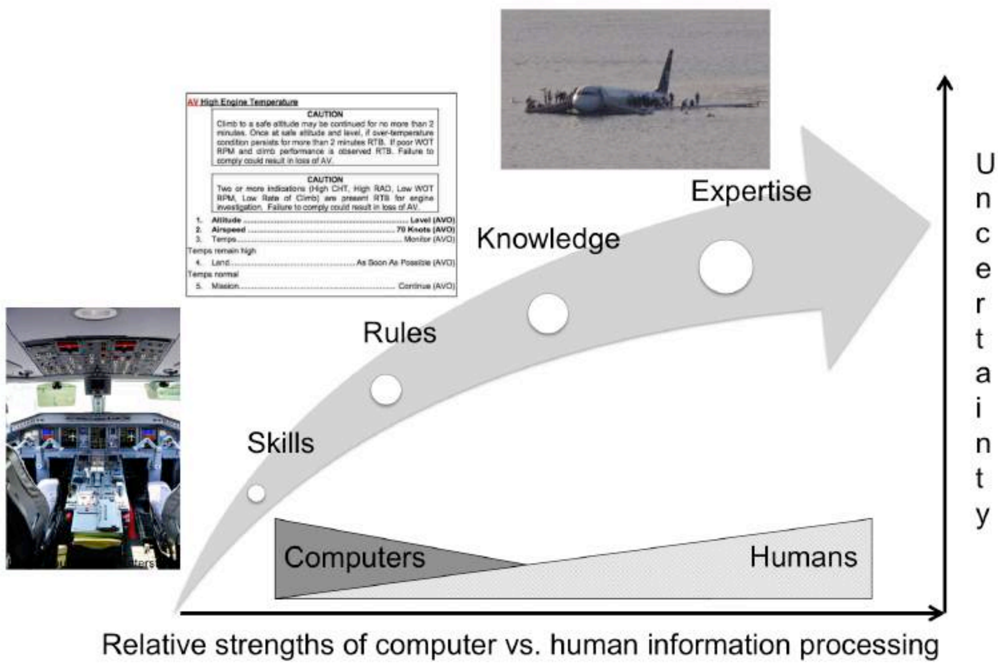

Figure 2: Skill, Rule, Knowledge, & Expert Behaviors and the Relationship to Uncertainty

by subroutines, stored rules, or procedures. For example, when a fire light illuminates or some other sub-system indicates a problem, pilots recognize that they should consult a manual to determine the correct procedure (since there are far too many procedures to be committed to memory), and then follow the steps to completion. Some interpretation is required, particularly for multiple system problems, which is common during a catastrophic failure like the loss of thrust in one engine. Recognizing which procedure to follow is not always obvious, particularly in warning systems where one aural alert can indicate different failure modes.

By the very nature of their if-then-else structures, rule-based behaviors are potentially good candidates for automation but uncertainty management is key, as depicted in Figure 2 on the vertical axis. If there is low uncertainty in the world, computer rule-based reasoning is typically superior than humans such as the ability of a car's navigation system to find the quickest path in a much faster time than humans. However, if there is traffic congestion or roadwork unknown to the automation, which are sources of uncertainty, human path planning is often superior.

This example highlights a conundrum of automation in the face of uncertainty. While fast and able to handle complex computation far better than humans, computer algorithms, which work primarily at the rule-based level, are notoriously brittle in that they can only take into account those quantifiable variables identified in the design stages that were deemed to be critical (Smith, McCoy, & Layton, 1997). In complex systems with inherent uncertainties (weather, traffic delays), it is not possible to include *a priori* every single variable that could impact the final solution.

Another problem for automation of rule-based behaviors is similar to one for humans, which is the selection of the right rule or procedure for a given set of stimuli. Computers will reliably execute a procedure more consistently than any human, but the assumption is that the computer selects the correct procedure, which is highly dependent on correct sensing. It is this reliance on near-perfect sensing that can make autonomous car rule-based reasoning extremely brittle, especially if the sensors do not work correctly. For example, poor computer vision was a major factor in the death of a Tesla driver who relied on the car's autopilot for obstacle detection (NTSB, 2017). The car could not execute its automated braking rule set, which generally is excellent, because it never "saw" a tractor-trailer crossing the road ahead and thus was never triggered.

While the ability of autonomous car perceptual sensors to map the world in a reliable and repeatable manner under all expected driving conditions is still very much an open engineering question, another relatively new area of concern for these perception systems is hacking. Even if autonomous car sensors work correctly under all weather conditions and in all traffic configurations, they are still not guaranteed to work correctly, especially if individuals with expertise know how to "trick" such systems.

Figure 3 depicts a stop sign that was "hacked" by researchers placing four black and white stickers in strategic locations on the sign (Evtimov et al., 2017), such that the computer

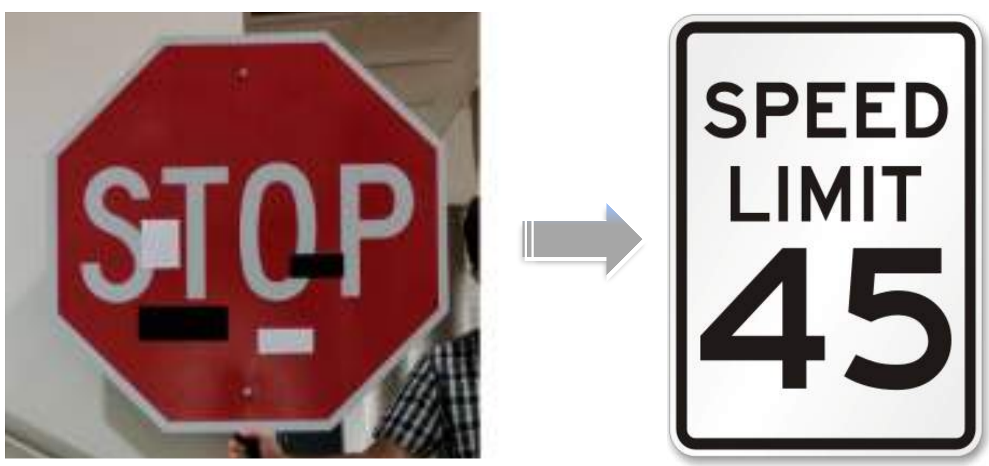

Figure 3: Using black and white stickers and a general attack algorithm called Robust Physical Perturbations, researchers can cause a computer vision system to see the stop sign as a 45 mph speed limit sign (Evtimov et al., 2017).

vision system, enabled through a deep learning network, "saw" the sign as a 45mph sign instead of a stop sign. If this were to happen in the real world, the car could select its increase-speed-to-45-mph algorithm instead of its stopping algorithm. While such demonstrations currently occur in research settings, it is inevitable that they will occur in the real world.

This example highlights that as uncertainty grows, both humans and computers will be required to reason inductively, as opposed to deductively because not all information can be known and estimates will have to be made. This is illustrated in Figure 2 as knowledge-based reasoning. Humans resolve these scenarios through judgment and intuition, and often rely on past experience to resolve such situations. This was the case in the landing of USAIR 1549 in 2009 in the Hudson River, as depicted in Figure 2. The Captain had to decide whether to ditch the aircraft or attempt to land it at a nearby airport. Given his mental model, the environment, and the state of the aircraft, his knowledge-based reasoning made him choose the ditching option, which led to a successful outcome.

The last level in the SRKE taxonomy is expertise, in which knowledge-based behaviors are a prerequisite for gaining expertise, but cannot be achieved without significant experience under the highest levels of uncertainty, which is difficult to replicate in test environments. Because of the aforementioned brittleness problems in computer algorithms and the inability to replicate the intangible concepts of judgment and intuition, knowledge-based reasoning and true expertise, for now, are outside the realm of computers. Advances are being made in areas of machine learning, including deep learning, which may help computers approximate something like knowledge-based reasoning. However, as demonstrated by the stop sign example in Figure 3, advancements in these areas are still very preliminary and significant research is needed to determine more reliable and robust methods for autonomous systems to reason under uncertainty.

Implications of Human Licensing on Autonomous Vehicle Certification

Given current licensing and certification processes of humans in both driving and aviation domains, and understanding the SRKE framework, what then can be learned from human licensing that could be applied to autonomous systems? Given that AVs will be increasingly introduced on the roadways, both federal and state agencies will quickly need answers to certification and licensing questions, which will also be increasingly needed in future drone operations.

Vision tests for AVs?

One important parallel that could be drawn between human and autonomous systems licensing is the very first test of any operator certification, the physical screening stage. More specifically, it is likely that licensing and certification tests of new autonomous systems will need to consider just how well these vehicles "see" the world around them.

As demonstrated in Figure 3, computer vision systems can fail miserably in what humans would see as a very simply inductive reasoning task. And such failures are not limited to intentional forms of hacking. Snow on top of signs, long shadows falling across signs in the late afternoon, graffiti, and foliage obscuring part of a sign or roadway elements are just some of the issues that can cause computer vision systems to not "see" correctly. Sensor washout in the early morning or late afternoon is another problem as well as faulty obstacle detections, such as the inability of many vision systems to distinguish writing in the sides of trucks from static signs.

How then can licensing agencies ensure that autonomous vehicles can "see", particularly in known areas of difficulty? Will new driverless cars be taken to test tracks and given partially obscured signs in the late afternoon or in snowing conditions to ensure that they operate as intended? Are there a set of images that can be "shown" to a computer vision system quickly, much like the Ishihara Color Plate Test where numbers made up of dots of one color are embedded in a field of other colors to ensure human can discriminate colors? Such types of tests that can quickly detect anomalies in software systems powered by probabilistic reasoning would likely be part of a larger multi-stage approach to testing, but as in all systems, early identification of significant issues is central to faster and less expensive development processes.

Knowledge tests and checkrides for AVs?

In terms of a written knowledge test, there are likely no direct certification parallels for the licensing of autonomous. However, just as there is industry consensus of what knowledge all human operators should have in present-day flying and driving domains (i.e., Tables 1 and 4), there should be a similar set of guidelines developed by NHTSA, who is responsible for regulating automobiles, that dictates the core body of "knowledge" that all Level 4 and 5 cars should have before they are safe to operate under various conditions. Once this minimum set of knowledge "requirements" are elucidated, then tests can be constructed either in simulation or in special test track environments that allow manufacturers to demonstrate that their cars meet such requirements.

These tests of knowledge, either in the real world or in simulation, then begin to approximate what humans experience as checkrides. Given the low level skill-based nature of driving practical exams which most autonomous cars today could easily pass (although their ability to follow verbal instructions would likely be abysmal), practical autonomous driving tests should model the checkride format of pilot licensing exams, especially the risk mitigation elements. Figure 2 highlights that as uncertainty and risks grow, the demand for knowledge and expert based reasoning also grows. Therefore, autonomous vehicle certification tests should focus on how the underlying software and hardware components perform under conditions of high uncertainty that could significantly jeopardize passenger safety as well as those individuals outside the car.

So if we start to combine the information in Table 1 with the format of an FAA licensing checkride (Table 4), then a set of requirements would emerge that dictate that AVs should be able to, for example, identify, assess and mitigate risks for positioning the car while crossing/entering and intersection with vision obstructed or identify, assess and mitigate risks when driving through shared work zones. Evaluators would then look for evidence from the manufacturers that models of AVs with various software and hardware combinations could indeed detect they are in one of these high risks areas of operations, and then develop an accurate world model of the environment, with *good enough* estimates of the potential risks and outcomes so that a safe course of action is generated by the car.

This emphasis on good enough is rooted in the decision theoretic concept known as satisficing (Simon, 1982), in which humans cannot have perfect knowledge of the world and so will generate solutions that are at least good enough to meet minimum constraints. The notion of licensing someone who performs not necessarily perfectly is well established in aviation. A critical element of these checkrides is that the examiner is attempting to assess the likelihood that an examinee will be able to perform correctly under even the highest levels of uncertainty. While there are clear standards for how checkrides should be administered, ultimately the examiner must make a judgment as to whether he or she *feels* the examinee is a safe *enough* operator. Thus there is an element of trust that is developed throughout the oral exam and checkride, and while the examinations are rooted in objective criteria, there are clearly subjective elements as well.

It is generally accepted that for AVs to be widely accepted, they will need to generate equivalent or better levels of safety (EBLS) than those of human-driven cars (Fraade-Blanar & Kalra, 2017). Thus autonomous vehicles will effectively need to be assessed on their ability to not necessarily behave perfectly in all settings, but to satisfice in a way that is superior to humans, especially in low probability but high consequence settings. Formally defining EBLS so that they can be assessed through either formal testing or naturalistic studies is still a step manufacturers and regulators have yet to address.

This need for regulators and evaluators to understand if and how AVs can generate plans and actions that meet EBLS is especially difficult given the opaque nature of the underlying machine learning algorithms, which are notoriously difficult to understand (Miller, Howe, & Sonenberg). The complexity of such approaches coupled with human difficulties in reasoning under uncertainty (Tversky & Kahneman, 1974) have motivated a relatively new field of inquiry into explainable or interpretable artificial intelligence (AI). Significant work is needed to develop more transparent algorithms, visualizations, and interaction modalities such that people with advanced degrees in machine learning can understand the limits of such approaches.

Graduated Licensing

In the United States and in many other countries like Australia and the United Kingdom, teenage drivers are certified through a staged process called graduated licensing (GL), although these programs vary in requirements. In GL programs, teen drivers are typically first allowed to drive only when supervised, and this license is often called a learner’s permit. This stage can be followed by a provisional licensing period where drivers are allowed to operate a vehicle unsupervised, but are subject to many restrictions such as limited hours of operation (e.g. daylight only), number and ages of passengers in the car, no cell phone in the car, etc. When experience and/or age requirements are met, the teen driver can obtain a full license.

Autonomous vehicles have arguably been undergoing a testing period similar to the learner’s permit process, with many states requiring a safety driver in the car to monitor the system and take over as necessary. However, in California AVs are now being allowed to operate on public roads with no human driver in the car, but with an operator remotely monitoring the vehicle’s progress and intervening as necessary through teleoperation. While such operations will undoubtedly yield valuable test data, they also need to be carefully considered. Relying on remote teleoperation to take over a vehicle in the case of an urgent or emergent scenario is extremely dangerous when required response times are on the order of seconds, which is typical of driving scenarios, especially those on the highway.

Humans suffer from an immutable physical limitation known as the neuromuscular lag, which is the ~ 0.5s delay that exists between the time of stimulus onset and the ability to react (Jagacinski & Flach, 2003). This inherent time delay is made worse in remote control of AVs given both communication delays and time required for remote operators to gain situation awareness in an information-impoverished remote environment. Indeed at one point in time, the US Air Force lost one third of its Predator unmanned aerial vehicle fleet because of the inability of remote operators to successfully control these vehicles, especially on landing which occurs at speeds similar to those of highway speeds (Erwin, 2009). It is highly unlikely that at highway speeds, a remote AV operator can effectively intervene to prevent an accident.

Given both the perceptual vision problems of AVs as outlined previously, as well as known problems with teleoperation of vehicles in time-critical scenarios, AVs should be held to a similar GL process with very clear restrictions on permissible areas, times, speeds, and weather conditions until such time that the AVs can demonstrate reliable operation under those conditions. Moreover, in other countries like England, Australia and India, inexperienced drivers are required to put highly visible signs on their cars with letters like L (learner) and P (provisional). The point of these signs is to heighten the awareness of nearby drivers to potentially unsafe behaviors, so they can adjust accordingly. Going forward, AVs should be required to have such clear markings, not only to alert human drivers of potential problems but also because in these early days which clearly constitute testing, human drivers should be made explicitly aware that they are taking part in a test (Cummings, 2017).

Certifying Machine Learning Algorithms is Unprecedented

One important difference between the licensing of human-operated and autonomous vehicles is the lack of individual differences. In both driving and flying, individuals are certified as safe to operate their vehicles and in the more complex case of flying, the practical exam can be tailored to various individuals as an examiner sees fit. Given that all models of a specific car or drone will presumably be identical, autonomous vehicles of the future will not need to be certified individually. Instead, they will likely need to be certified in groups and classes.

For example, various manufacturers have already proposed different driverless car designs, which means that groups of cars will have different hardware and software solutions between manufacturers but it is also possible that manufacturers will have different solutions on different models. This raises the question of component testing and certification, i.e., if a LIDAR is certified on one model of a Ford car, are all LIDARs on all their cars certified? What if cars use the same hardware, but different software post-processing algorithms?

Software upgrades are another issue that will need to be addressed in licensing and certification. Tesla is the first car company to introduce automatic over-the-air software updates for its cars, but as self-driving cars come online with potentially unreliable capabilities, will major upgrades that introduce new capabilities need certification before pushed to owners who will not likely read the owner's manuals and any associated warnings?

Software bugs and failures are commonplace in current cars, without any advanced autonomy or embedded machine learning. Failures to detect software bugs were implicated in the Toyota unintended acceleration cases leading to a 1.2B settlement with the US government (Douglas & Fletcher, 2014). Software bugs in general were responsible for 15% of recalls in 2015, and the number of software-related components involved in recalls grew by almost 600% between 2011-2015 (Steinkamp, 2016). These problems occurred despite car companies having access to well-established test practices for traditional software. Given that there are no established test standards for probabilistic reasoning software like machine learning, this rate of software errors with resulting recalls will likely significantly increase.

The automotive industry will likely need to model how the FAA certifies new software upgrades for fly-by-wire aircraft, which share many commonalities with AVs. The FAA relies on an office in the Aviation Safety branch, the Aircraft Certification Service, for approval of software and airborne electronic hardware for airborne systems including autopilots and flight controls, which have direct parallels in the automotive industry. One key aspect of the FAA's software certification process is their partnership with a not-for-profit association called the RTCA (Radio Technical Commission for Aeronautics). This group works with the FAA and the aviation industry to develop consensus-driven performance standards and guidance materials that serve as a partial basis for certification of critical systems and equipment used in the conduct of air transportation.

This certification of complex hardware and software aviation systems through consensus has been generally successful, but there are examples of incomplete or faulty certifications such as the Boeing 787 Battery fire in 2013, where the FAA's certification process was called into question (NTSB, 2014). However, as software grows in complexity, it is not clear that the consensus approach to autonomous system certification is going to be sufficient. Current testing approaches for such systems rely heavily on deterministic testing, meaning that for every set of inputs, the system will act or react the same way within some small confidence interval. Such testing processes are highly repeatable and for both hardware and software components in the air and on the road today, there are widely accepted industry test standards,

Autonomous systems are stochastic, as opposed to deterministic ones, which means they rely on probabilistic reasoning and significant estimation. They heavily leverage machine learning, aka deep learning, which is a data-intensive approach to developing an autonomous system world model, which serves as the core set of assumptions about who, what and where agents in the system are and what their likely next set of behaviors and actions will be. To date, there exists no industry consensus on how to test such systems, particularly in safety-critical environments of high uncertainty. Given this lack expertise, the FAA's approach to software certification may serve as a starting point for future autonomous vehicle certifications, but significantly work is needed in academic and industry to address a significant knowledge gap in certifying control algorithms that rely on imperfect sensors and data-driven reasoning to make decisions in safety-critical systems. As mentioned previously, significant more work is needed in the emerging field of Explainable AI in order to fill this knowledge gap.

Conclusion

Currently there is significant legislation pending at both state and federal levels to introduce autonomous vehicles into the marketplace. Regardless of the outcomes, autonomous vehicles will be introduced into the everyday driving landscape, but significant questions remain as to how such vehicles could or should be certified as safe enough to operate at equivalent or better levels of safety demonstrated by human drivers.

In both driving and aviation settings, humans, who are autonomous agents, are certified through a three tiered process which focuses on physical readiness, demonstration of broad system knowledge of operation, and practical exams, aka checkrides, where operators are tested on their skills and in the case of aviation, their ability to mitigate risk in the face of uncertainty. While such certification processes will not be needed for humans in the case of Level 4 and Level 5 vehicles, there is still much that can be learned from how humans are certified to safely operate vehicles and aircraft and then applied to autonomous vehicles of all types.

First, while AVs will not need to be assessed on their physical readiness per se, their vision systems contain many weaknesses, with more emerging in basic research demonstrations almost daily. Such weaknesses are due to both sensor limitations but also software-driven post-processing which makes them vulnerable to hacking. Given that these perception systems are the heart of any autonomous car, it is critical that weaknesses in these systems are both known and mitigated, making testing of these systems especially critical.

Second, the body of knowledge that human drivers are tested on is highly relevant when determining EBLOS of AVs. Moreover, just as the FAA requires human pilots to exhibit demonstrable knowledge across various areas of operations including how to mitigate risk for known contingencies, AVs should be required to demonstrate their boundaries of competence in driving scenarios that range from mundane to life threatening. More specifically, given the known weaknesses of perception systems, it is particularly important that both engineers and regulators have a clear understanding of how the probabilistic algorithms embedded in AVs perceive and mitigate risk. Unfortunately, research into explainable AI is just beginning and so the limits of such approaches is still not well understood.

It is important to point out that these arguments and comparisons are primarily targeting the certification of Level 4 and 5 vehicles, which assume the driver is not relied upon for intervention. The licensing/certification of Level 3 vehicles contains elements of both human licensing and vehicle certification, which complicates which agencies would be ultimately responsible since NHTSA typically is responsible for vehicle safety while state agencies are responsible for driver licensing.

Level 3 vehicle certification is very difficult and beyond the scope of this chapter since such operations require coordinating the handing of vehicle control back and forth between the computer and human. Because human interaction with autonomous systems is plagued with known serious issues such as mode confusion (Bredereke & Lankenau, 2002), boredom and complacency (Cummings & Gao, 2016), and distractions and delayed reaction times (Ranney, 2008), licensing and certification are not separable issues. In these cases, one small vehicle design change can have dramatic effects on driver behavior and thus licensing and certification should be seen as an integrated effort as opposed to mutually exclusive efforts.

Self-driving and driverless cars carry great promise both in terms of reduced driver deaths and increased accessibility for drivers with a myriad of disabilities or a lack of access to a personal vehicle. However, those technologies that will enable these achievements are very immature, with no industry consensus of a minimum set of safety standards, or how such cars should be tested including identification of the corner cases that define worse possible scenarios. Significantly more work is needed to develop principled testing protocols and there are many important lessons to be learned from how humans have been licensed in driving and aviation domains.

One critical advancement that is needed for both engineers and regulators of future autonomous systems of all types is the development of explainable AI methods that give both engineers and regulators insight into both how such probabilistic systems achieve their solutions and when such solutions are brittle and likely to fail. While accidents and deaths ultimately lead to safer systems, especially for emerging technologies (Petroski, 1992), we can and should identify major weaknesses before deadly problems materialize in the population at large.

Acknowledgements

This work was inspired by Dr. Michael Francis and supported in part by the US Department of Transportation and the University of North Carolina's Collaborative Sciences Center for Road Safety (CSCRS).

References

American Association of Motor Vehicle Administrators. (2014). *AAMVA GUIDELINES for Noncommercial Knowledge and Skills Test Development*. Retrieved from Arlington, VA:

American Medical Association. (2003). Chapter 9: Medical Conditions and Medications That May Impair Driving *Physician's Guide to Assessing and Counseling Older Drivers*: National Highway Transportation and Safety Administration,.

Bredereke, J., & Lankenau, A. (2002). *A Rigorous View of Mode Confusion.* Paper presented at the SafeComp, Bremen, Germany.

Cummings, M. L. (2014). Man vs. Machine or Man + Machine? *IEEE Intelligent Systems,, 29*(5), 62-69.

Cummings, M. L. (2017). The Brave New World of Driverless Cars,: The Need for Interdisciplinary Researchnand Workforce Development. *TR News*, 34-37.

Cummings, M. L., & Gao, F. (2016). Boredom in the Workplace: A New Look at an Old Problem. *Human Factors, 58*(2), 279-300.

Cummings, M. L., Stimpson, A., & Clamann, M. (2016). *Functional Requirements for Onboard Intelligent Automation in Single Pilot Operations.* Paper presented at the AIAA SciTech Conference, San Diego, CA.

Douglas, D., & Fletcher, M. A. (2014). Toyota reaches $1.2 billion settlement to end probe of accelerator problems. *Washington Post*

Erwin, S. (2009). UAV Programs Illustrate DoD's Broken Procurement System. *National Defense*

Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., . . . Song, D. (2017). Robust Physical-World Attacks on Deep Learning Models. *arXiv preprint 1707.08945*.

Flight Standards Service. (2017). *Commercial Pilot – Airplane Airman Certification Standards*. (FAA-S-ACS-7 (Changes 1 & 2)). Washington DC: U.S. Department of Transportation,.

Fraade-Blanar, L., & Kalra, N. (2017). *Autonomous Vehicles and Federal Safety Standards: An Exemption to the Rule?* Retrieved from Santa Monica, CA:

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

Miller, T., Howe, P., & Sonenberg, L. (2017). *Explainable AI: Beware of Inmates Running the Asylum*. Paper presented at the IJCAI-17 Workshop on Explainable AI (XAI) Proceedings Melbourne, Australia.

National Center for Statistics and Analysis. (2017). *Traffice Safety Facts: Research Note*. Washington DC: Department of Transportation,.

Nneji, V., Stimpson, A., Cummings, M. L., & Goodrich, K. (2017). *Exploring Concepts of Operations for On-Demand Passenger Air Transportation,.* Paper presented at the AIAA Aviation, Denver, CO.

NTSB. (2014). *Auxiliary Power Unit Battery Fire Japan Airlines Boeing 787-8, JA829J Boston, Massachusetts January 7, 2013*. (NTSB/AIR-14/01 PB2014-108867). Washington DC: National Transportation Safety Board,.

NTSB. (2017). *Highway Accident Report: Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida May 7, 2016*. (NTSB/HAR-17/02 PB2017-102600). Washington DC.

Petroski, H. (1992). *To Engineer is Human*. New York: Vintage Books.
Ranney, T. A. (2008). *Driver Distraction: A Review of the Current State-of-Knowledge*. Washington DC: Department of Transportation
Rasmussen, J. (1983). Skills, rules, and knowledge: Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics, 13*(3), 257-266.
Shinar D, & Schieber F. (1991). Visual requirements for safety and mobility of older drivers. *Human Factors., 33*(5), 507-519.
Simon, H. A. (1982). *Models of Bounded Rationality* (Vol. 2). Cambridge, MA: MIT Press.
Smith, P. J., McCoy, C. E., & Layton, C. (1997). *Brittleness in the design of cooperative problem-solving systems: The effects on user performance.* Paper presented at the IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans.
Steinkamp, N. (2016). *2016 Automotive Warranty & Recall Report*. Retrieved from
Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.
Virginia Department of Motor Vehicles. (2017). *Virginia Driver's Manual*. Commonwealth of Virginia,.
Wiegmann, D. A., & Shappell, S. A. (2001). *A Human Error Analysis of Commercial Aviation Accidents Using the Human Factors Analysis and Classification System (HFACS)*. Washington DC: Federal Aviation Administration.