Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings

Mary L. Cummings

■ *Artificial intelligence, in the form of machine learning, has the potential to transform many safety-critical applications such as those in transportation and healthcare. However, despite significant investment and impressive demonstrations, such technologies have struggled to live up to their promises. To this end, this article illustrates that machine learning fundamentally lacks the ability to leverage top-down reasoning, a critical element in safety-critical systems. This is especially important in situations where uncertainty can grow very quickly, requiring adaption to unknowns. This fundamental lack of contextual reasoning, combined with a lack of understanding of what constitutes maturity in artificial intelligence-embedded systems, has significantly contributed to the failures of these systems. Demonstrations where safety-critical artificial intelligence-enabled systems function as if they were almost operational should not be a substitute for testing. Instead, companies and regulatory agencies need to work together to develop clear criteria and certification protocols before such technologies are made publicly available.*

While artificial intelligence (AI) has recently been touted as very successful across a number of domains, including transportation, medical applications, and digital personal assistants, the reality that such systems may not actually be as capable as envisioned is slowly creeping into the national consciousness. While AI can show up in many everyday applications from shopping to management of home automation, it is the application of AI in safety-critical systems such as transportation and medicine that is the most concerning — because, literally, the incorrect use of AI can have deadly consequences.

For example, in transportation settings, it has been well established that AI is unable to cope with unexpected poses of known objects; a motorcycle laying on the ground after an accident may not actually be seen *as* a motorcycle (Alcorn et al. 2018). Problems with automotive computer vision have been cited as contributing factors in many fatal Tesla crashes (Crowe 2016; Lohr 2016) and the death of a pedestrian in an Uber self-driving car accident (Griggs and Wakabayashi 2018). Despite years of promises by many

 ISSN 0738-4602

companies of full-self driving powered by AI, many companies have walked back their claims in attempt to recalibrate the public's and funders' expectations (Bubbers 2019; Elias 2019).

Another major area where AI has been heralded as a success is in healthcare settings, including drug discovery (Morrison 2019) and radiology (Ardila et al. 2019; Park et al. 2019). While these successes are important steps toward making AI a useful tool in aiding diagnostic applications, there have also been many spectacular failures. IBM's Watson, the decision-making engine behind the *Jeopardy!* AI success, has been deemed a costly and potentially deadly failure when extended to medical applications such as cancer diagnosis (Strickland 2019). Alphabet's DeepMind medical AI applications are facing similar questions (Lu 2019).

In concert with public backlash over AI and privacy, as well as concerns with AI embedded in social media that could be manipulating people, negative sentiment is growing about applications of AI. Many experts are concerned that this backlash could lead to another AI winter, which could lead to significant distrust in legitimate AI advances and a cooling of financial support (Walch 2019). Given this potential outcome, it is important to step back and analyze just why AI is struggling to gain traction in safety-critical systems and how the roadmap to success would need to change to achieve positive outcomes.

To this end, this article will first argue that, in current formulations, AI that leverages machine learning (ML) fundamentally lacks the ability to leverage top-down reasoning, which is a critical element in safety-critical systems where uncertainty can grow very quickly requiring adaption to unknowns. Then, this article will explain how this lack of fundamental understanding combined with a lack of understanding of what constitutes maturity in AI-embedded systems has contributed to the potential failure of these systems. This article concludes with recommendations for human-AI collaborative systems as well as paths forward to mitigate the impact of AI misapplications and better inform future uses.

The Problem of Brittleness

In safety-critical settings such as transportation and healthcare, computer vision is a common application of AI, which typically means algorithms leverage machine-, sometimes called deep-, learning to *perceive* the world to make decisions. For example, deep learning algorithms in driverless cars determine whether a car *sees* a pedestrian; or in healthcare, whether a tumor exists in a grainy image of a lung. While important advancements have been made in the last ten years in computer vision and in the deep learning algorithms that underpin these systems, such approaches to developing perceptual models of the real world are plagued by problems of brittleness.

Brittleness occurs when any algorithm cannot generalize or adapt to conditions outside a narrow set of assumptions. For example, many natural language processing algorithms are brittle when they can understand a person from New York City but fail to understand the same sentence from someone in Appalachia or who speaks English with a foreign accent (Harwell 2018). While this brittleness may be frustrating for a person attempting to navigate a phone tree, it can be deadly in a safety-critical system that relies on any kind of ML for perception or critical reasoning.

The source of this perceptual brittleness comes from the fact that ML algorithms do not actually learn to perceive the world in a way that can generalize in the face of uncertainty. For example, computer vision ML algorithms typically rely on edge detection to decompose an image through mathematical computations to identify transitions between dark and light colors. These transition points then become a set of line segments, hence the term *edges*. Figure 1 is an example of how a picture can be decomposed into its edges. So, while humans see a tiger, a deep learning algorithm sees sets of lines in various clusters.

For an ML algorithm to learn to recognize a tiger, it must see tens of thousands of similar images to understand patterns of reoccurrence. Such patterns ideally scale spatially so that even potentially at different distances and angles, the object can be successfully detected. What the algorithm has learned is that a particular set of mathematical relationships belong together as a label for a particular object. Once an algorithm can classify an object correctly, it can invoke a set of rules for how to treat that object; for example, if one is in a car and a tiger (or any other animal) is in the car's path, then the car should stop.

Algorithm brittleness occurs when the environment changes in such a way that the computer vision algorithm can no longer recognize the object due to some small perturbation. Recognizing animals like tigers in images has been dramatically improving due to ML research, but images with multiple species and unusual behaviors can cause problems for identification (Norouzzadeh et al. 2017).

Such problems are also seen in safety-critical settings like driving. Brittleness for driverless car computer vision includes an inability to cope with changes caused by weather conditions. Lane markings that are partially covered by snow cause problems because the edges no longer match the system's internal model (Krishner 2019). Even on sunny days, when a tree branch or other vegetation partially obscures just a traffic sign, what is obvious to a human becomes impossible to interpret for a computer vision algorithm (Lewis 2019).

A common response to such brittleness is for engineers and computer scientists to gather more data to fill what is thought to be a perceptual gap. For example, to fix the vegetation-obscuring-a-sign problem, many engineers will say "We just need more examples to train the algorithm to correctly recognize this condition." While that is one answer, it begs the questions

Figure 1. Edge Decomposition Example.

Courtesy of Wessam Bahnassi.

as to how much of this finger-in-the-dyke engineering is practical or even possible? Every time a new sensor is created (like a new light detection and ranging [LIDAR] sensor) and every time this sensor experiences a new set of conditions it has not yet seen, it must be trained with a significant amount of data that may have to be collected. The workload to do this is extremely high, which is one reason why there is such a talent drain caused by the current driverless car space race. All this intense effort, which has a significant cost, is occurring for systems with significant vulnerabilities.

Because computer vision based on deep learning is still a relatively new area of research, new problems are coming to light in university laboratories. Researchers have only recently uncovered that neural nets are not capturing accurate depth information in images (van Dijk and de Croon 2019), which can have significant safety implications. A relatively new field of study has emerged in the past few years called adversarial ML, which examines how systems that leverage versions of deep learning algorithms can be tricked or defeated.

Progress in adversarial ML has been eye-opening, as one set of researchers demonstrated that putting four innocuous black and white stickers on a stop sign could trick a computer vision algorithm to see a forty-five miles-per-hour speed limit sign (Evtimov et al. 2017). Another set of researchers then went on to show that only a single pixel needed to be changed to cause such an algorithm to mislabel an object (Su, Vargas, and Sakurai 2019). These recent efforts show just how vulnerable these ML-based approaches are in computer vision applications, and, ultimately, how nascent this field still is.

Bottom-Up versus Top-Down Reasoning

A fundamental issue with ML algorithm brittleness is the notion of bottom-up versus top-down reasoning, which is a basic cognitive science construct. It is theorized that when humans process information about the world around them, they use two basic approaches to making sense of the world: bottom-up and top-down reasoning. Bottom-up reasoning occurs when information is taken in at the sensor level — the eyes, the ears, the sense of touch, and so forth — to build a model of the world. Top-down reasoning occurs when perception is driven by cognition expectations. These two forms of reasoning are not mutually exclusive as humans use their sensors

to gather information about the world, but often apply top-down reasoning to fill in the gap for information that may not be known.

For example, in the case of a lane marking covered by snow, humans leverage bottom-up reasoning to see the snow and partially covered lines and then use top-down reasoning to infer where the line would be even if they can't see all of it. Humans do not need perfect information in the world because of their ability to fill in missing information from experience and abstract reasoning happens almost instantly with little-to-no previous experience (and certainly not requiring tens of thousands of examples to make such an educated guess).

The human ability to infer relationships from partial information is captured visually by the Kanizsa triangle in figure 2. A computer vision algorithm would learn that this image has three equally spaced Vs with three alternately spaced circles, each missing a one-sixth piece. Such deconstruction is effectively bottom-up reasoning. However, because of experience and expectation, most humans will see two triangles superimposed over one another, an example of top-down reasoning. While the label *triangle* could be assigned to this image (by a human programmer) as well as thousands of other similar images in an attempt to teach a computer what such abstractions mean, — up to this point in time, ML algorithms have been unsuccessful at both recognizing and creating visual illusions (Williams and Yampolskiy 2018).

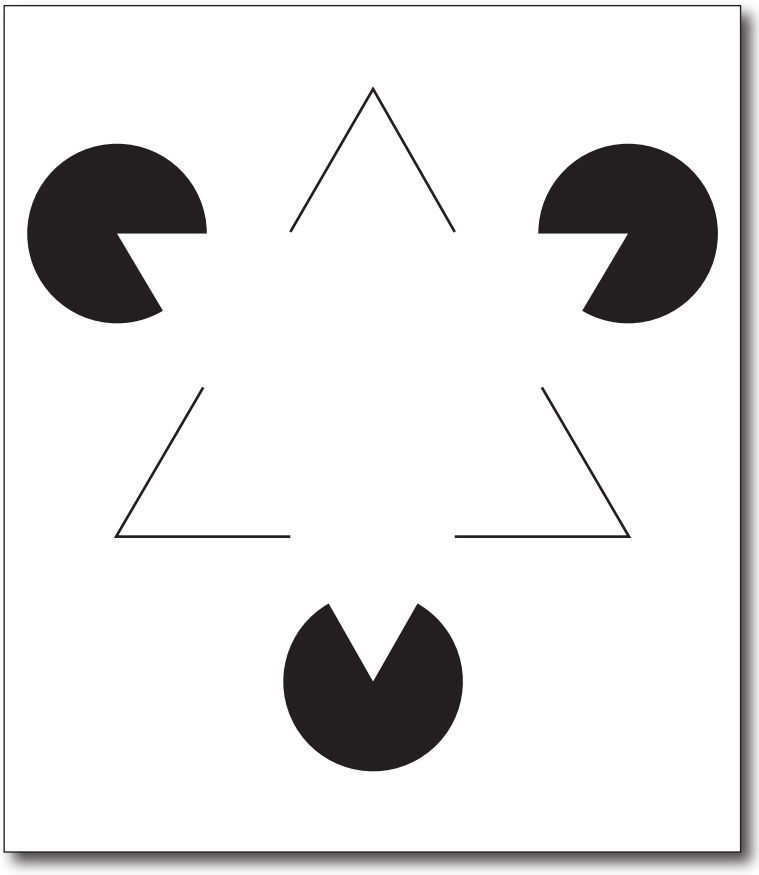

Figure 2. The Kanizsa Triangle Visual Illusion.

While there has been success of using ML in limited vision contexts, such algorithms effectively only apply only half of the reasoning needed to solve complex problems — the bottom-up construction of individual pieces of data. Deep learning algorithms are quite shallow in that they can detect patterns, but they lack any sense of causality — which is critical for understanding what to do in novel situations.

What is missing is the ability to consistently consider context as well as lower the uncertainty due to missing or degraded pieces of information, which are the missing pieces that knowledge- and expert-based reasoning apply. The lack of ML top-down reasoning in perceptual tasks is why computer vision algorithms struggle with labeling unexpected images in transportation settings (Alcorn et al. 2018) and fail in radiology applications (Hosny et al. 2018).

To illustrate how and why both bottom-up and top-down reasoning is needed for complex decision-making, figure 3 depicts the kinds of reasoning needed for such a task, independent of who (the human and/or the computer) performs it. This skills-rules-knowledge-expertise (SRKE) depiction (Cummings 2014) is an extension of Rasmussen's skills, rules, and knowledge-based behaviors taxonomy (Rasmussen 1983).

Skill-based behaviors are the lowest point in the taxonomy and consist of sensory-motor actions that are highly automatic and typically acquired after some period of training (Rasmussen 1983). In figure 3, an example of skill-based control for humans is the act of keeping a car within lane lines, which easily becomes a highly automated skill once learned. If uncertainty is low at this stage, for example, all the information is available for how to do a particular task, such reasoning is an ideal candidate for automation. Indeed, automated lane keeping is now a standard feature on many cars.

Once a set of basic skills is acquired, like those in driving between two lane lines, operators can then turn their attention to higher cognitive tasks such as rule-based behaviors, which are effectively those actions guided by subroutines, stored rules, or procedures. For example, when a driver (or a computer) detects a stop sign, a set of procedures that leverage various skills are involved, like slowing the car down and bringing the car to a stop before the sign. As depicted in figure 3, uncertainty is somewhat higher at this stage, primarily due to the need to infer which set of stored rules or procedures is needed at a particular time or place.

The next level in the SRKE taxonomy is that of knowledge-based behaviors, where mental models built over time aid in the formulation and selection of plans for an explicit goal (Rasmussen 1983). Those scenarios where knowledge-based reasoning is needed are typically characterized by higher uncertainty.

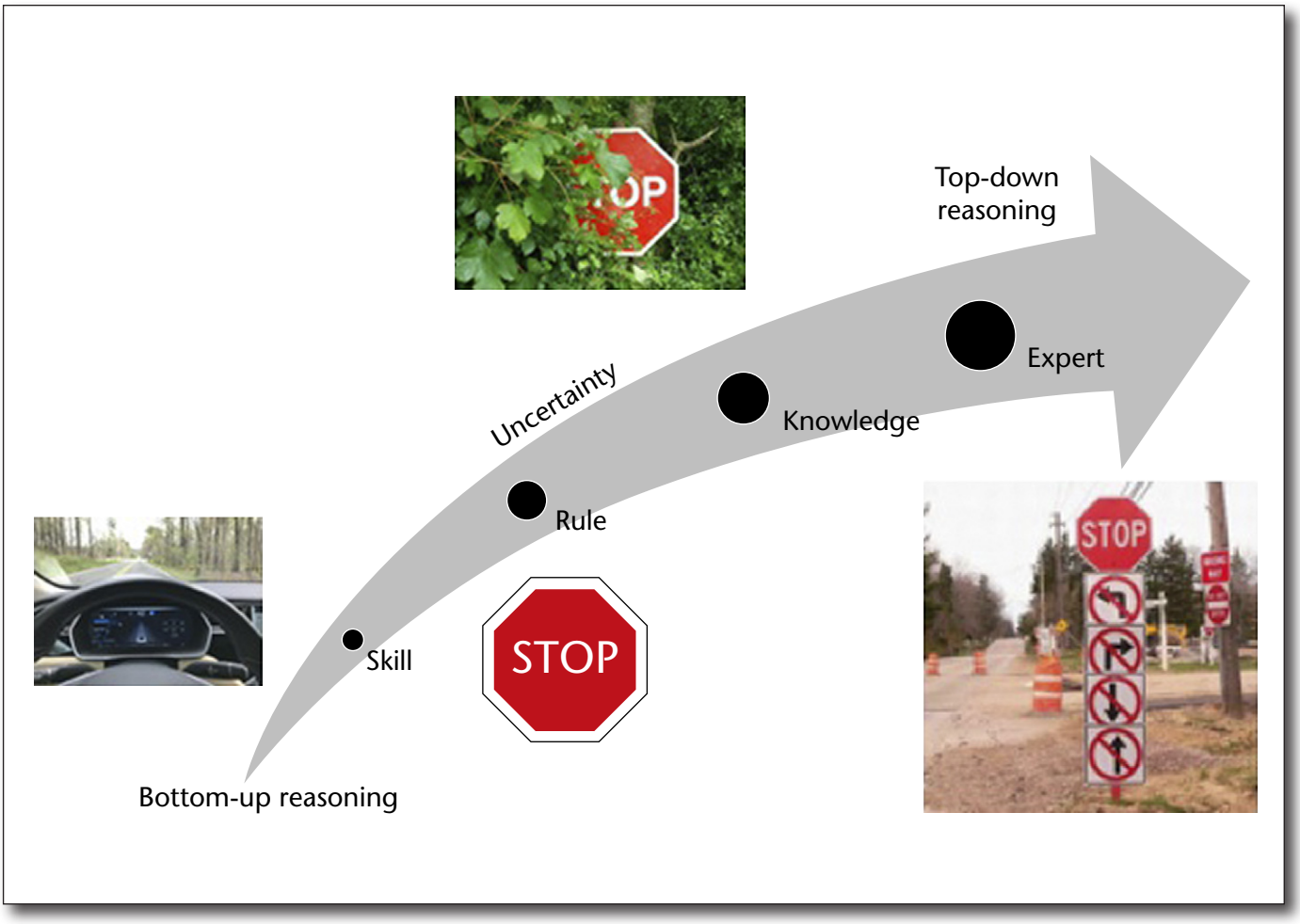

Figure 3. Bottom-Up and Top-Down Reasoning in Light of the SRKE-Based Behavior Taxonomy.

Occluded stop sign courtesy of Albert Bridge, road sign near Donegore, County Antrim, Northern Ireland. CC BY-SA 2.0.

In figure 3, human drivers leverage knowledge-based reasoning when they see, for example, a partially occluded sign like a stop sign covered by vegetation. The entire sign is not visible, but faced with this uncertainty, drivers easily surmise that it is a stop sign and then they know to invoke the required rule-based reasoning as a result.

The last behavior in the SRKE taxonomy is expertise. Figure 3 demonstrates that expertise must be leveraged under the highest levels of uncertainty, where decision-makers find themselves in situations that cannot precisely be determined, with potentially many unknown variables. Judgment and intuition are the key expert behaviors that allow for quick assessment of uncertain situations, typically in a fast and frugal manner (Gigerenzer and Todd 1999). The expert-reasoning scenario of multiple conflicting traffic signs in figure 3 demonstrates an extremely confusing and uncertain scenario, with no clear set of rules to rely upon, requiring significant judgment to resolve.

As depicted in figure 3, skill-based reasoning requires significant bottom-up processing of information and it is at this stage of information processing where ML-enabled computers perform well, assuming the sensors can accurately and reliably obtain exogenous information. So, for situations where skill-based reasoning dominates and sensors can reliably develop world models, ML-enabled systems can perform quite well. Indeed, many companies have demonstrated impressive self-driving scenarios under sunny conditions and well-marked roads, which is primarily due to low uncertainty and the ability to stay on the lower end of the SRKE spectrum.

However, while the bulk of driving may reside at the low end of the SRKE taxonomy as pictured in figure 3, there are occasions that require top-down reasoning that computers simply are currently not equipped to solve. Recently a driverless shuttle was involved in a crash because it could not understand the intent of a human driver of a tractor-trailer ahead of it who very slowly backed up for more

maneuvering room, expecting the shuttle to also back up (NTSB 2019). The tractor-trailer driver did not know the shuttle had no driver (nor did the shuttle have the ability to operate in reverse).

The driver had an expectation built over years of experience that the other vehicle would give way and be able to reverse, but the shuttle had no rule set to reference. This scenario seems simple for human drivers who understand the need to negotiate to resolve uncertainty, but such abstract principles and the development of alternative action plans, even simple ones, is outside of the realm of ML-enabled systems, at least for the foreseeable future.

Such ambiguous situations happen regularly in the driving domain and often with much more dramatic and deadly consequences. There have been several incidents where Tesla drivers have been killed while driving on Autopilot, an automated driving assist feature, which failed to see objects directly in cars' paths, and a pedestrian was killed by an Uber self-driving car while undergoing human-supervised testing (Crowe 2016; Griggs and Wakabayashi 2018; Lohr 2016). In all these cases, the skill-based reasoning automated systems that relied on bottom-up processing failed, and deaths occurred because the inattentive drivers did not realize these cars still needed their top-down reasoning and judgment.

These examples highlight the essential need for any safety-critical system to incorporate both bottom-up and top-down reasoning, especially as uncertainty grows in a system. This is true whether such uncertainty is caused by confusing scenarios in the external environment or by failures in the sensors to build accurate world models. Unfortunately, because of the nascent nature of ML-enabled technologies and the hypercompetitive nature of Silicon Valley, it is not always obvious to the engineers developing these technologies that their creations may not adequately reason across the spectrum as pictured in figure 3, and are too immature for deployment.

The next section will discuss how companies in the past have known whether their technologies were mature enough for deployment and what milestones should be achieved before fielding a technology with embedded ML in an operational setting. Most start-ups and other Silicon Valley-based companies pride themselves on working differently and faster than traditional companies, but the cost of this speed and agility is that many important lessons that more traditional companies have learned over the years, may be missed.

Not All Demonstrations Are Equal

To allow various programs across the National Aeronautics and Space Administration the ability to accurately gauge the abilities of new proposed technologies in the space program, in the 1970s the Technology Readiness Levels (TRLs) framework was proposed. Originally seven and now nine, as seen in figure 4, the nine TRLs qualitatively describe where a potential technology sits in relation to its maturity and likelihood of readiness for deployment (Hirshorn 2017). This framework allows people to evaluate technologies through a shared language, and has been adopted for use across the US Department of Defense (DOD 2017), the US Department of Energy (DOE 2010), the US Federal Highway Administration (Towery, Machek, and Thomas 2017), and many others.

When originally conceptualized, the TRL levels focused on primarily physical systems that were predominantly leveraging new hardware developments. The words *model*, *prototype*, and *component* suggest a physical item that can been touched and seen. Even the term "breadboard" refers to a physical circuits-and-electronics board where initial designs were conceptualized. Curiously, the word *software* never shows up in any of the TRL levels, despite the increasing prevalence of software in such complex systems.

The US government has been broadly criticized for its lack of understanding of the importance of software development and how a lack of explicit consideration in the systems acquisition process can lead to long and costly delays (McQuade et al. 2019). While it is well recognized that the US government needs to overhaul its software engineering practices, what is less clear is how the lack of understanding of software maturity complicates the overall TRL framework in figure 4. Indeed, immaturity in both software testing and acquisition processes has been cited as major causes of delays in the US Department of Defense F-35 aircraft program. The number of extensive and costly delays in the program after it was deemed to have reached TRL 9, which is operational capability, suggest serious mistakes were made in assessing whether the whole system, including the software, was actually mature enough for operations (GAO 2019).

The military is not the only entity to suffer from lapses in accurately assessing the readiness of new technology. In the civilian aviation world, the recent Boeing 737 MAX groundings are an example of what happens when immature and untested software code is embedded in an aircraft thought to be a physically mature platform. Because versions of the Boeing 737 (a TRL 9 platform) have been flying for well over fifty years, there was a cavalier assumption that the software code did not have to be treated as a new "component" for an aircraft with such a long history of physical implementation. The 737 MAX control software was nothing like that of older aircraft, probably at a TRL of 5 to 6, and not at all ready for operational deployment at TRL 9. Given its flight criticality, even though the airframe was thought to a be a more mature technology, the entire system's TRL was only as good as its lowest common denominator.

What lessons, if any, could be learned from the government's mistakes in developing new technologies that are thought to be mature, but do not account for the immaturity of embedded software? One study of thirty-seven US Department of Defense

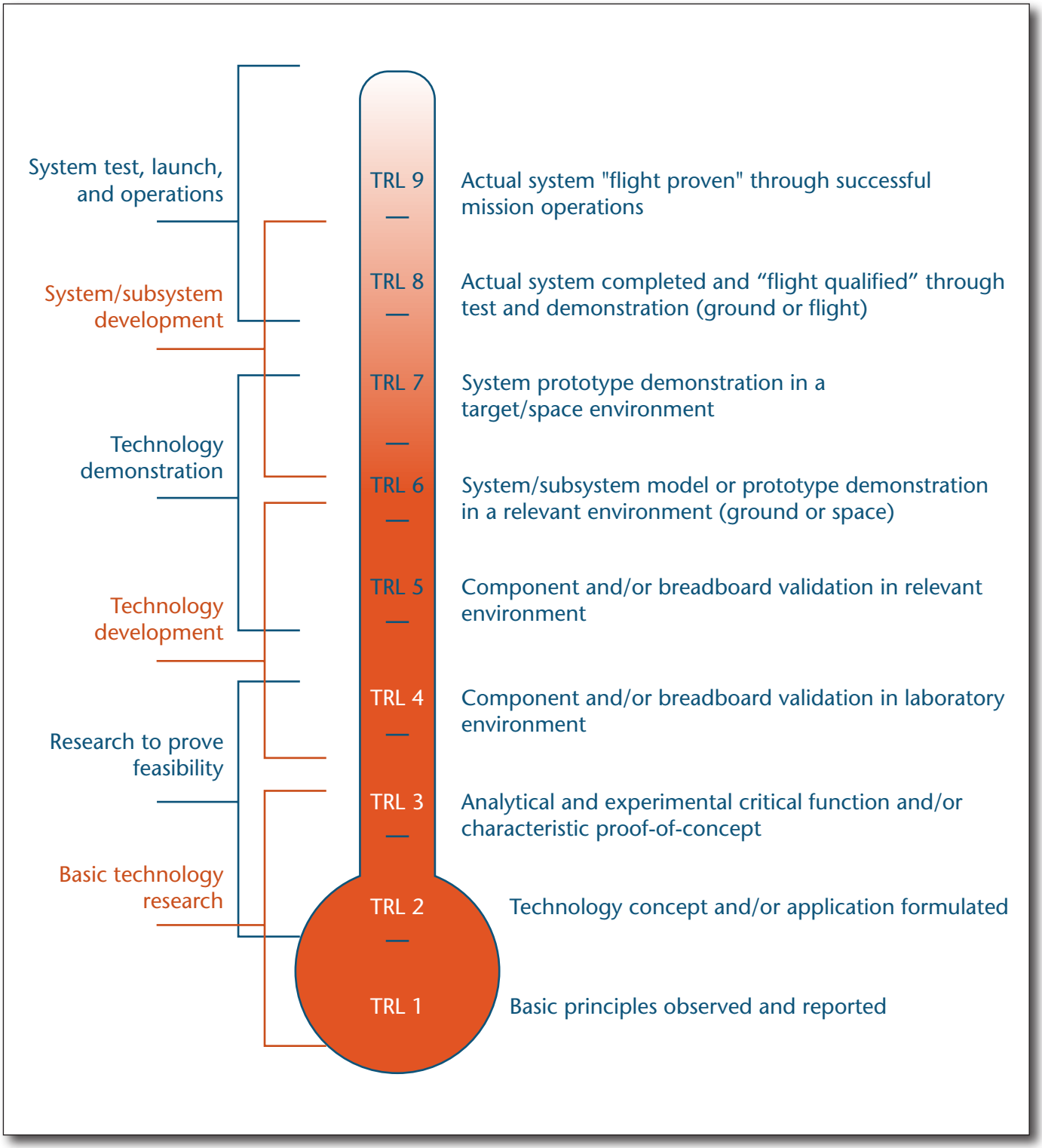

Figure 4. National Aeronautics and Space Administration TRLs.

weapon systems showed that a lack of technology maturity understanding had a statistically significant effect on schedule overruns (Katz et al. 2014). The Government Accounting Office has stated that risk is acceptably low for product development for systems at or above TRL 7 when there has been a "demonstration of a technology in its final form, fit, and function within a realistic environment" (GAO 1999, 2001). In terms of autonomous systems that incorporate significant layers of AI, like in driverless cars,

it is critical to further examine what it means to be in "its final form, fit, and function within a realistic environment."

When a technology is in its final form, one would expect that not only are the hardware elements fairly stable, but that the software code underpinning the perception, sensor fusion, and control algorithms has also reached some measure of stability. It is not clear, in the case of self-driving cars, that either hardware or software maturity has been reached. There is broad consensus across the self-driving car industry that LIDAR sensors are critical for safe operations, but the LIDAR industry is still in significant flux and many new types and kinds of LIDARs have recently been introduced (Lienert and Klayman 2019).

Changing or significantly upgrading a LIDAR has a direct impact on software stability in that all sensor fusion algorithms require recalibration and retraining whenever a new sensor is inserted into the hardware *stack* (the perception and control system of an autonomous vehicle). So, if a critical hardware component is not yet stable, then it is impossible for the associated software to be stable. Moreover, LIDAR is not the only sensor expected to change in the near-term, as radars are also expected to undergo significant upgrades (Murray 2019) and new types of 3D cameras are making their way to the market (Dent 2019).

In addition to the Government Accounting Office's recommendation for a technology to be in its final form, the other important attribute worth consideration is what it means to perform in a realistic environment. To reach broad market appeal, self-driving cars will need to operate in all weather conditions and under different levels of road quality. Self-driving systems, even with their multiple sensors and software advancements, still cannot reliably work in rain and snow conditions (Zang et al. 2019), during time of low sun angles (Dowling 2019), and often where lines on the road are either nonexistent or are present with faded paint (Sage 2016).

While many self-driving companies have produced impressive demonstrations in places like Arizona and California, such limited applications and the high number of conditions in which they cannot currently operate suggest that these technologies are actually at TRL 6, where a prototype demonstration has occurred in a relevant environment. Indeed, the biggest difference in whether a technology is TRL 6 or TRL 7 is performing in a relevant versus a realistic environment. This one seemingly nuanced difference is easy to overlook, but could have many unexpected consequences when missed.

The problems with asserting that a technology is TRL 7 when it is actually at a TRL 6, like that of driverless cars, can be quite dramatic. The Government Accounting Office looked at military technologies that were assumed to be TRL 7 when in fact they were TRL 6 and found that sixty percent of cost growth in programs occurred after the technology moved into production. Typically, these programs declared themselves to be production-ready and operational before fully completing testing in realistic (as opposed to relevant) settings, and then they ultimately failed (GAO 2017).

Rapidly moving products to production with embedded AI before they are ready has been a distinct trend for many Silicon Valley-backed technologies. While the Theranos debacle is often labeled as outright fraud, it is just an extreme example of the "fake-it-'til-you-make-it" Silicon Valley culture and elements of such an attitude has occurred across numerous application of AI through Wizard of Oz techniques where humans pretend to be AI (Solon 2018). The fake-it-'til-you-make-it attitude is simply recognition that a technology is something less than TRL 7 but is then advertised as more mature than it actually is.

One of the ramifications of such a fake-it-'til-you-make-it culture is inflated and unrealistic expectations that drive a hypercompetitive first-to-market race, which can become prohibitively expensive. In the case of the driverless car industry that has surpassed $100 billion in investment (Eisenstein 2018), it is not clear if the industry can withstand a sixty percent or more cost growth as it moves into the production phase with a significant risk of failure, just like the military programs with similar pedigrees of claiming to be more mature than they actually are.

Conclusion

AI, in the form of ML, has the potential to transform elements of many safety-critical applications and offers up new forms of human–computer collaboration that previously were out of reach. For example, one military-sponsored project recently demonstrated that an AI-enabled robotic arm could assist the pilot of an airplane in nonessential mundane tasks (Aurora Flight Sciences 2016). This is especially important because there is currently a global pilot shortage and so this kind of human augmentation could free copilots to take captain roles and effectively double the workforce. In a related medical example, many believe that the power of AI in radiology is not in the replacement of doctors but in assisting them in triaging images (Liew 2018).

Although AI augmentation of humans in safety-critical systems is well within reach, this success should not be mistaken for the ability of AI to replace humans in such systems. Such a step is exponential in difficulty and with the inability of ML, or really any form of AI reasoning, to replicate top-down reasoning to resolve uncertainty, AI-enabled systems should not be operating in safety critical systems without significant human oversight.

To address the known gaps in the brittleness of AI, there has been recent increasing interest in the fusing of symbolic and connectionist approaches to AI. Symbolic AI, the more classic form of AI, attempts to represent abstract human knowledge through the encoding of facts and rules (that is, symbols), and is commonly used in expert systems. Deep Blue, the

IBM chess-playing computer that outwitted Garry Kasparov, is an example of symbolic AI. AI in the form of ML is a connectionist AI approach, which loosely mimics neural connections in the brain in the form of probabilistic networks that represent information and simulated intelligence (Marcus and Davis 2019; Toews 2019). An AI algorithm that detects cancerous nodes in radiologic images, based on its training of thousands of images with such cancers previously labeled, is an example of connectionist AI.

Unfortunately, the fusing of symbolic and connectionist AI will not fundamentally solve the brittleness problem from which both approaches suffer, nor will fusing the two have any ability to solve the top-down reasoning issue. As per figure 3, connectionist AI approximates bottom-up reasoning and symbolic approaches represent rule-based reasoning, with some overlap between the two. Neither approach can handle significant uncertainty, and neither (or even both together) can approximate top-down reasoning, problems with context, and the need for judgment under uncertainty. Real breakthroughs in AI will not be achieved until some form of contextual and casual-based computational approach is developed.

Even though AI has limits, particularly in safety-critical systems with potentially deadly latent conditions, demanding perfection could limit the benefits of developing such technology. As in the case of the robot pilot arm or in the case of slow-speed driverless shuttles that operate in protected environments, there may be very advantageous uses of such AI-enabled systems, even though the technology is not flawless. This then motivates the need to develop clear criteria and testing protocols so that companies and governments buying or approving AI-enabled systems can be sure that the proposed systems are indeed at TRL 7 and capable of operating in their intended operational domains.

However well-intended, companies that demonstrate that their AI-enabled systems, especially those that operate in safety-critical settings, can *almost* function as if they were operational is simply not a high enough bar. History is replete with examples of how similar promises of operational readiness ended in costly system failure, and these cases should serve as a cautionary tale to not just the driverless car community, but to all the AI researchers and practitioners that subscribe to the "fake-it-'til-you make-it" mantra.

References

Alcorn, M. A.; Li, Q.; Gong, Z.; Wang, C.; Mair, L.; Ku, W. S.; and Nguyen, A. 2018. Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. arXiv preprint arXiv:1811.11553 [cs.CV]. Ithaca, NY: Cornell University Library.

Ardila, D.; Kiraly, A.P.; Bharadwaj, S.; and Choi, B. 2019. End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography. *Nature Medicine* 25: 954–61. doi.org/10.1038/s41591-019-0447-x.

Aurora Flight Sciences. 2016. Aurora Demonstrates DARPA Aircraft Autonomy Program, October 17. Manassas, VA: Aurora Flight Sciences.

Bubbers, M. 2019. Don't Hold Your Breath — Fully Autonomous Cars Are Still Decades Away. The Globe and Mail, May 27.

Crowe, S. 2016. Tesla Autopilot Causes 2 More Accidents. Robotics Trends, July 12. www.roboticsbusinessreview.com/rbr/tesla_autopilot_causes_2_more_accidents/.

Cummings, M. L. 2014. Man vs. Machine or Man + Machine? *IEEE Intelligent Systems* 29(5): 62–9. doi.org/10.1109/MIS.2014.87.

Dent, S. 2019. Outsight's 3D Camera for Autonomous Cars Can Identify Clothing and Ice. Engadget, September 17. www.engadget.com/2019/09/17/outsight-cedric-hutchings-3d-self-driving-camera.

DOD. 2017. *Defense Acquisition Guidebook*. K. Stewart, editor. Washington, DC: Defense Acquisition University, Department of Defense (DOD).

DOE. 2010. *Standard Review Plan: Technology Readiness Assessment Report*. Edited by Office of Environmental Management. Washington, DC: US Department of Energy (DOE).

Dowling, B. 2019. Self-Driving Cars in Boston Blinded by Solar Glare at Traffic Lights. Xconomy, February 20. xconomy.com/boston/2019/02/20/nutonomy-self-driving-autonomous-vehicles-boston-sun-glare-traffic-light.

Eisenstein, P. A. 2018. Not Everyone Is Ready to Ride as Autonomous Vehicles Take to the Road in Ever-Increasing Numbers. CNBC, October 14. www.cnbc.com/2018/10/14/self-driving-cars-take-to-the-road-but-not-everyone-is-ready-to-ride.html.

Elias, J. 2019. Alphabet Exec Says Self-Driving Cars Have Gone Through a Lot of Hype, but Google Helped Drive That Hype. *CNBC*, October 23. www.cnbc.com/2019/10/23/alphabet-exec-admits-google-overhyped-self-driving-cars.html.

Evtimov, I.; Eykholt, K.; Fernandes, E.; Kohno, T.; Li, B.; Prakash, A.; Rahmati, A.; and Song, D. 2017. Robust Physical-World Attacks on Deep Learning Models. arXiv preprint arXiv:1707.08945 [cs.CR]. Ithaca, NY: Cornell University Library.

GAO. 1999. *Better Management of Technology Development Can Improve Weapon System Outcomes*. Washington, DC: US Government Accounting Office (GAO).

GAO. 2001. *Best Practices: Better Matching of Needs and Resources Will Lead to Better Weapon System Outcomes*. Washington, DC: US Government Accounting Office (GAO).

GAO. 2017. *Defense Acquisitions: Assessments of Selected Weapon Programs*. Washington, DC: US Government Accountability Office (GAO).

GAO. 2019. *F-35 Aircraft Sustainment: DOD Needs to Address Substantial Supply Chain Challenges*. Washington, DC: US Government Accounting Office (GAO).

Gigerenzer, G., and Todd, P. M. 1999. *Simple Heuristics That Make Us Smart*. S. Stich, editor. Oxford, UK: Oxford University Press.

Griggs, T., and Wakabayashi, D. 2018. How a Self-Driving Uber Killed a Pedestrian in Arizona. The New York Times, March 21. www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html.

Harwell, D. 2018. The Accent Gap. The Washington Post, July 19. www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

Hirshorn, S. R. 2017. *NASA Systems Engineering Handbook, Revision 2*. Edited by Aeronautics Research Mission Directorate.

Washington, DC: National Aeronautics and Space Administration (NASA).

Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L. H.; and Aerts, H. 2018. Artificial Intelligence in Radiology. *Nature Reviews. Cancer* 18(8): 500–10. doi.org/10.1038/s41568-018-0016-5.

Katz, R.; Sarkani, S.; Mazzuchi, T.; and Conrow, E. H. 2014. The Relationship of Technology and Design Maturity to DoD Weapon System Cost Change and Schedule Change During Engineering and Manufacturing Development. *Systems Engineering* 18(1): 1–15. doi.org/10.1111/sys.21281.

Krishner, T. 2019. 5 Reasons Why Autonomous Cars Aren't Coming Anytime Soon. AP News, February 4. apnews.com/article/b67a0d6b6413406fb4121553cdf0b95a.

Lewis, R. K. 2019. Reality Is Going to Stall for Some Time the Advent of Driverless Cars. The Washington Post, August 3. www.washingtonpost.com/realestate/reality-is-going-to-stall-for-some-time-the-advent-of-driverless-cars/2019/08/01/343c9458-afa8-11e9-a0c9-6d2d7818f3da_story.html.

Lienert, P., and Klayman, B. 2019. A Chaotic Market for One Sensor Stalls Self-Driving Cars. Reuters, March 5. www.reuters.com/article/us-autos-autonomous-lidar-focus/a-chaotic-market-for-one-sensor-stalls-self-driving-cars-idUSKCN1QN0HW.

Liew, C. 2018. The Future of Radiology Augmented with Artificial Intelligence: A Strategy for Success. *European Journal of Radiology* 102: 152–6. doi.org/10.1016/j.ejrad.2018.03.019.

Lohr, S. 2016. Blind Spots Ahead. The New York Times, September 20, Section D, 1. https://www.nytimes.com/2016/09/20/science/computer-vision-tesla-driverless-cars.html.

Lu, D. 2019. It's Too Soon to Tell if DeepMind's Medical AI Will Save Any Lives. New Scientist, July 31. www.newscientist.com/article/2212100-its-too-soon-to-tell-if-deepminds-medical-ai-will-save-any-lives/.

Marcus, G., and Davis, E. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon.

McQuade, J. M.; Murray, R. M.; Louie, G.; Medin, M.; Pahlka, J.; and Stephens, T. 2019. Software Acquisition and Practices (SWAP) Study. Software Is Never Done: Refactoring the Acquisition Code for Competitive Advantage. Defense Innovation Board. Washington, DC: Office of the Secretary of Defense. https://media.defense.gov/2019/Mar/14/2002101480/-1/-1/0/DIB-SWAP_STUDY_REPORT[DRAFT],_LAST%20MODIFIED_13MAR2019.PDF.

Morrison, C. 2019. AI Developers Tout Revolution, Drugmakers Talk Evolution. Nature Biotechnology, November 8. www.x-mol.com/paper/5938356. doi: 0.1038/d41587-019-00033-4.

Murray, C. 2019. Autonomous Cars Look to Sensor Advancements in 2019. Design News, January 7. www.designnews.com/electronics-test/autonomous-cars-look-sensor-advancements-2019/95504860759958.

Norouzzadeh, M. S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M.; Packer, C.; and Clune, J. 2017. Automatically Identifying, Counting, and Describing Wild Animals in Camera-Trap Images with Deep Learning. arXiv preprint arXiv:1703.05830 [cs.CV]. Ithaca, NY: Cornell University Library.

NTSB. 2019. Low-Speed Collision Between Truck-Tractor and Autonomous Shuttle, Las Vegas, Nevada, November 8, 2017. Washington, DC: National Transportation Safety Board (NTSB).

Park, A.; Chute, C.; Rajpurkar, P.; Lou, J.; Ball, R. L.; Shpanskaya, K.; Jabarkheel, R.; Kim, L. H.; McKenna, E.; Tseng, J., Ni, J.; Wishah, F.; Wittber, F.; Hong, D.S.; Wilson, T. J.; Halabi, S.; Basu, S.; Patel, B. N.; Lungren, M. P.; Ng, A. Y.; and Yeom, K. W. 2019. Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. *JAMA Network Open* 2(6): e195600. doi.org/10.1001/jamanetworkopen.2019.5600.

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* SMC13(3): 257–66. doi.org/10.1109/TSMC.1983.6313160.

Sage, A. 2016. Where's the Lane? Self-Driving Cars Confused by Shabby U.S. Roadways. Reuters, March 31. www.reuters.com/article/us-autos-autonomous-infrastructure-insig/wheres-the-lane-self-driving-cars-confused-by-shabby-u-s-roadways-idUSKCN0WX131.

Solon, O. 2018. The Rise of "Pseudo-AI": How Tech Firms Quietly Use Humans to Do Bots' Work. The Guardian, July 6. www.theguardian.com/technology/2018/jul/06/artificial-intelligence-ai-humans-bots-tech-companies.

Strickland, E. 2019. IBM Watson, Heal Thyself: How IBM Overpromised and Underdelivered on AI Health Care. *IEEE Spectrum* 56(4): 24–31. doi.org/10.1109/MSPEC.2019.8678513.

Su, J.; Vargas, D. V.; and Sakurai, K. 2019. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation* 23(5): 828–41. doi.org/10.1109/TEVC.2019.2890858.

Toews, R. 2019. To Understand the Future of AI, Study Its Past. Forbes, November 17. www.forbes.com/sites/robtoews/2019/11/17/to-understand-the-future-of-ai-study-its-past/#6888673821b3.

Towery, N. D.; Machek, E.; and Thomas, A. 2017. *Technology Readiness Level Guidebook*. J. A. Volpe, editor. Cambridge, MA: US Department of Transportation.

van Dijk, T., and de Croon, G. C. H. E. 2019. How Do Neural Networks See Depth in Single Images? arXiv preprint arXiv:1905.07005 [cs.CV]. Ithaca, NY: Cornell University Library. doi.org/10.1109/ICCV.2019.00227.

Walch, K. 2019. Are We Heading for Another AI Winter Soon? Forbes *Cognitive World*, October 20. www.forbes.com/sites/cognitiveworld/2019/10/20/are-we-heading-for-another-ai-winter-soon/#5347c1f256d6.

Williams, R. M., and Yampolskiy, R. V. 2018. Optical Illusions Images Dataset. arXiv preprint arXiv:1810.00415 [cs.CV]. Ithaca, NY: Cornell University Library.

Zang, S.; Ding, M.; Smith, D.; Tyler, P.; Rakotoarivelo, T.; and Kaafar, M. A. 2019. The Impact of Adverse Weather Conditions on Autonomous Vehicles: How Rain, Snow, Fog, and Hail Affect the Performance of a Self-Driving Car. *IEEE Vehicular Technology Magazine* 14(2): 103–11. doi.org/10.1109/MVT.2019.2892497.

Mary L. Cummings is a professor in the Electrical and Computer Engineering Department at Duke University and is the director of the Humans and Autonomy Laboratory. Her current research focuses on human supervisory control, explainable AI, human-autonomous system collaboration, human-robot interaction, human-systems engineering, and the ethical and social impact of technology.