



MARJORY S. BLUMENTHAL, LAURA FRAADE-BLANAR, RYAN BEST, J. LUKE IRWIN

Safe Enough

Approaches to Assessing Acceptable Safety for
Automated Vehicles



For more information on this publication, visit www.rand.org/t/RAA569-1

Library of Congress Cataloging-in-Publication Data is available for this publication.

ISBN: 978-1-9774-0603-3

Published by the RAND Corporation, Santa Monica, Calif.

© Copyright 2020 RAND Corporation

RAND® is a registered trademark.

Cover: elenabs//Adobe Stock

Limited Print and Electronic Distribution Rights

This document and trademark(s) contained herein are protected by law. This representation of RAND intellectual property is provided for noncommercial use only. Unauthorized posting of this publication online is prohibited. Permission is given to duplicate this document for personal use only, as long as it is unaltered and complete. Permission is required from RAND to reproduce, or reuse in another form, any of its research documents for commercial use. For information on reprint and linking permissions, please visit www.rand.org/pubs/permissions.

The RAND Corporation is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest.

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Support RAND

Make a tax-deductible charitable contribution at
www.rand.org/giving/contribute

www.rand.org

Preface

Stakeholders in the field of automated vehicles (AVs)—from developers to the general public—seek confidence in the safety of these machines. In summer 2019, the Uber Advanced Technologies Group approached the RAND Corporation for help in understanding (a) approaches for establishing that AVs are acceptably safe, and (b) which approaches appeal more and less broadly across different categories of stakeholders. The analysis described in this report builds on material that the Uber Advanced Technologies Group requested of RAND in summer 2017; that research examined how to develop a vendor- and technology-neutral framework for measuring AV safety that could be suitable for broad use, including among developers, and culminated in the 2018 RAND report *Measuring Automated Vehicle Safety: Forging a Framework*.¹ In our research for both reports, we drew heavily on consultations with participants in the AV industry, industry observers from government, and safety-focused researchers and nonprofit organizations. For this report, we also had the benefit of a novel survey of the general public, complementing comments collected through consultations with other kinds of stakeholders about how the public thinks.

Community Health and Environmental Policy Program

RAND Social and Economic Well-Being is a division of the RAND Corporation that seeks to actively improve the health and social and economic well-being of populations and communities throughout the world. This research was conducted in the Community Health and Environmental Policy Program within RAND Social and Economic Well-Being. The program focuses on such topics as infrastructure, science and technology, community design, community health promotion, migration and population dynamics, transportation, energy, and climate and the environment, as well as other policy concerns that are influenced by the natural and built environment, technology, and community organizations and institutions that affect well-being. For more information, email chep@rand.org.

¹ Laura Fraade-Blanar, Marjory S. Blumenthal, James M. Anderson, and Nidhi Kalra, *Measuring Automated Vehicle Safety: Forging a Framework*, Santa Monica, Calif.: RAND Corporation, RR-2662, 2018.

Contents

Preface	iii
Boxes, Figures, and Tables.....	vi
Summary.....	vii
Acknowledgments	xii
Abbreviations	xiii
1. Introduction	1
Defining Safety Is Problematic	1
AV Safety in Context	4
“Acceptably Safe” and This Report	5
2. Safety as a Codified Concept for AVs	8
3. Safety as a Measurement	11
How Measures Show Safety: Lagging and Leading Measures.....	11
Validity and Feasibility of Measures for Assessing Safety.....	22
Uniformity Versus Customization of Safety Measures.....	23
4. Safety as a Process.....	26
Safety Cases as Crosscutting Presentations of Evidence for Acceptably Safe	27
Safety as Indicated by Compliance with Technical Standards or Best Practices.....	28
Safety as Indicated by Compliance with Federal, State, or Local Regulations.....	31
Safety as Indicated by Corporate Safety Culture	33
5. Safety as a Threshold.....	37
Safety as Achieving a Threshold Based on Human Driving Performance	39
Safety as Achieving a Threshold Based on ADS Technology Potential.....	46
Safety as Achieving a Threshold Based on an Absolute Safety Goal.....	48
How Thresholds Can Be Used	52
6. Communicating About Safety	57
Risk Perception	57
Attitudes Toward Technology.....	61
Convincing the Public	64
Communication Among Stakeholders	68
Appraisal	69
7. Conclusions and Recommendations	71
Conclusions	71
Mapping Agreement and Disagreement.....	77
Stakeholder Ecosystem	78
Improving Communication About AV Safety	81
Next Steps	83

Appendix A. American Life Panel Survey	86
Appendix B. Interviews	94
Appendix C. Literature Highlights	97
Bibliography	110

Boxes, Figures, and Tables

Boxes

Box 1.1. Summary of <i>Measuring Automated Vehicle Safety: Forging a Framework</i>	2
Box 5.1. ALARA and ALARP	55

Figures

Figure S.1. Approach Relationships	ix
Figure 2.1. Approach Relationships	9
Figure 3.1. Illustrating Selected Driving Events as a Continuum	12
Figure 7.1. How Approaches Come Together to Show Evidence of and Support Communication About Acceptable Safety of AVs	73
Figure 7.2. How Approaches Build on and Provide Evidence in Support of Each Other	74
Figure 7.3. How Process and Measurement Approaches Build on and Support Each Other	75
Figure A.1. Sample Survey Prompt	88
Figure A.2. Perceived Safety Ratings, by Safety Message Source and Content	90

Tables

Table S.1. Approaches for Assessing AV Safety	viii
Table 2.1. Approaches for Assessing AV Safety	8
Table 3.1. Relationship Between Hard Braking and Appropriate ADS Action	15
Table 3.2. Strengths and Weaknesses of Selected Measures' Abilities to Reflect or Fail to Provide Evidence of Acceptable Safety	18
Table 5.1. Ability of Each Threshold Type to Provide Evidence of Acceptable Safety	53
Table 5.2. Applying ALARA and ALARP Phraseology to Approach	56
Table 6.1. Order of Explicit Rankings and Implicit Influence Attributed to Differing Sources for Messages About AV Safety	67
Table 7.1. Approaches to Assessing AV Safety	72
Table A.1. Linear Mixed Model Predicting Safety Ratings Using Safety Message Source, Age, and Gender	89
Table A.2. Means and Standard Deviations of Perceived Safety, by Survey Item	92

Summary

Automated vehicles (AVs) are coming to America’s roadways. They are not coming as quickly as was forecast five years ago—partly because the people developing them now have a clearer understanding of how difficult it is to make them safe—but incremental progress continues to be made in improving AV safety. This progress adds urgency to the need to understand when AVs can be considered *acceptably safe*—that is, safe enough to operate on public roads without the oversight of a human, professional safety driver.

In this report, we examine different approaches for appraising whether AVs are acceptably safe. Our analysis draws from three data sources: interviews with a diverse group of AV stakeholders, a survey of the general public, and a review of relevant literature. We also consider areas of agreement and disagreement among different groups of stakeholders about the value of different approaches. Finally, we examine the importance of communicating to public audiences about AV safety.

Approaches to Assessing AV Safety

We developed the following categorization of approaches for assessing AV safety:

- safety as a measurement
- safety as a process
- safety as a threshold.

Table S.1 summarizes the approaches we considered and the principal takeaways from our analysis about the strengths and weaknesses of each.

Safety as a measurement encompassed several specific methods of assessment within the broader category of measurement because AV technology is too immature to lend itself to the simple safety measurements that exist for conventional automobiles, such as *lagging measures* (actual counts of events). Developers are working on more-complex measurements (*leading measures*, which signal what lagging measures might eventually show) and seeing encouraging, if incomplete, results. One promising measurement method that assesses how an AV behaves in traffic is captured by the concept of *roadmanship* (the ability to drive on the road safely without creating hazards and responding well to hazards created by others) as introduced in RAND’s *Measuring Automated Vehicle Safety: Forging a Framework*.² Versions of this concept emerged in several of our interviews.

² Laura Fraade-Blonar, Marjory S. Blumenthal, James M. Anderson, and Nidhi Kalra, *Measuring Automated Vehicle Safety: Forging a Framework*, Santa Monica, Calif.: RAND Corporation, RR-2662, 2018.

Because existing measures are insufficient to assess AV safety, more attention is being paid to *processes*—the kinds of steps taken by developers and how these steps are implemented—and what they indicate about AV safety. These processes include safety-relevant standards-setting activities, a growing emphasis on safety culture, and widening use of *safety cases* (collections of assertions defining how safety is being promoted and assessed). Compliance with regulations is also a process, but its scope remains smaller than that for conventional vehicles. Most progress involving safety improvement processes emphasizes technical and managerial activity.

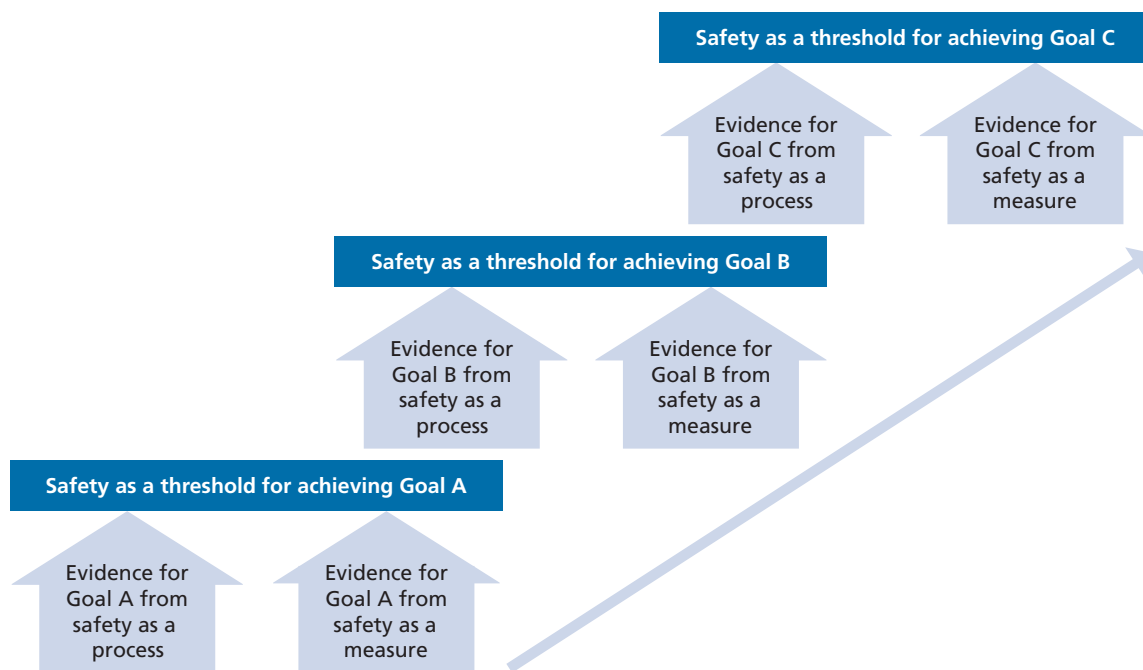
Finally, we examine *thresholds* as a way to assess safety. The most straightforward kind of threshold is the comparison between AVs and human drivers. People gravitate to comparisons with the familiar, but comparing human drivers with automated driving systems is harder than it might seem. The more experts consider such comparisons, the more they favor comparison with safe or better-than-average human drivers. Other kinds of thresholds relate to AV performance and to absolute goals, such as the Vision Zero goal (a global movement aimed at eliminating traffic fatalities by 2050). Thresholds will evolve as the technology develops, commercial usage expands, and expectations rise. Meeting thresholds is not a one-time achievement.

Table S.1. Approaches for Assessing AV Safety

Approach	Approach Takeaways
Measurement Leading measures Roadmanship Lagging measures	<ul style="list-style-type: none"> Measurement provides established and easily communicated evidence of safety, but its usefulness can be constrained by (a) a gap between what developers do internally and what can be shared externally for competitive reasons or concerns about how well past data reflect existing safety performance; (b) a lack of meaningful thresholds to contextualize measurement results, resulting in either a lack of comparisons or misleading ones; and (c) immaturity of leading measures, which continue to be developed along with the technology but remain works in progress. Roadmanship, a leading measure of roadway citizenship, is implicit in leading-measurement development today.
Process Technical standards Government regulation Safety culture	<ul style="list-style-type: none"> Given the imperfect fit of automated driving systems to regulations developed for conventional vehicles and the constraints on publicly shared, valid measurements, work to adapt existing and develop new processes provides indicators about the quantity and quality of developer attention to safety. Safety cases are a crosscutting process feature.
Thresholds Predicated on human driving Predicated on automated driving system technological performance Predicated on an absolute goal	<ul style="list-style-type: none"> Thresholds exist in qualitative and quantitative forms. They can be informed by measures or processes or considered on their own. Meeting thresholds is an ongoing, evolving process: There is not a one-time threshold; and thresholds can be internal, for developers, or external, for a broader group of stakeholders. The lure of comparing automated and human driving is strong. Difficulties arise with implementing human-driving and technology-based thresholds. Absolute thresholds might serve as a transportation-system-wide goal as opposed to an AV-specific one. Thresholds can evolve.

None of the approaches functions alone. Rather, these approaches complement, support, and interact with each other. Both processes and measures provide evidence for or against AVs achieving a level of acceptable safety. The approaches discussed can provide common scaffolding for both assessing AV safety and communicating about it. Figure S.1 illustrates how these components fit together. Each approach builds off of, supports, and contributes to the evidence used in others. For example, evidence from safety measures and processes can be used to show progress toward meeting a given threshold. Achieving AV safety is an ongoing effort; evidence of safety provides a staircase to achieving each threshold. Given the limitations to what can be measured or compared (as explained in this report), approaches should be used in combination.

Figure S.1. Approach Relationships



Communicating About AV Safety

Communicating about AV safety is essential: Stakeholders and the public must be assured that AVs are safe. Yet the continuing evolution of AVs and the fact that the technology is a black box to the public and stakeholders outside industry complicate this communication. Although messages about AV safety come from a variety of sources (government at different levels, AV developers, and safety researchers and advocates), the provenance of messages is not always obvious, at least to the general public. A survey conducted for this project through the RAND American Life Panel demonstrates the importance of how communications with the general

public are structured. Communication about AV safety should be data-driven but not too complex; it should speak to real-world scenarios; and it should come, in particular, from authoritative bodies without obvious bias (i.e., government at different levels).

Communication about AV safety must factor in people's difficulty in accurately gauging risk. AVs have emerged at a time when much is known about the challenges of communicating about risks and the psychology of how the general public hears messages about the risks of everyday products and experiences and weighs these risks. Communication is also important because, as with other technologies associated with significant risk, AV safety is a political issue and a technical one.

Mapping Agreement Among Stakeholders

Stakeholders interviewed for this project agreed on a few things:

- the value of data and statistics and an unwillingness to rely on expert opinion alone
- the inevitability of comparisons between automated driving systems and human drivers
- the criticality of company safety culture.

Although all kinds of stakeholders accept that 100-percent safety is never achievable, differences persist in how that issue is discussed and used as a basis for practical decisions. The disparity between what people in the industry know and what other kinds of stakeholders know (*information asymmetry*) shapes the politics of AV safety and leads to disagreements or gaps in understanding among different stakeholders.

Conclusions

AV development is proceeding as a commercial venture, funded by private companies. It is expected that a combination of superior safety (compared with conventional automobiles driven by people) and the potential for new business models will make the large investments pay off. To achieve this vision, consumer trust is essential. Agreeing on approaches to assessing AV safety and improving communication with consumers about AV safety are important for building and sustaining that trust.

Stakeholders expect AV safety to improve regularly, just like AV software does. The approaches presented and analyzed in this report can be useful to all kinds of stakeholders for gauging progress in AV safety.

Disagreement persists over when to start and scale up AV exposure on public roadways. There is no consensus about what approaches should be used to aid in such decisions. The disagreement is over how to choose the performance floor—which is, in effect, a selection of a threshold.

According to a survey done for this project, the public appears to have the greatest confidence in the government as a source of AV safety information. That finding should provide

motivation for the kinds of ongoing dialogue between the industry and government at different levels that both officials and the safety advocacy community exhort.

AV developers, perhaps through one of the coalitions that has emerged, should collaborate on creating publicly accessible (i.e., comprehensible to laypeople) versions of safety cases. Whereas the Voluntary Safety Self-Assessments encouraged by the U.S. Department of Transportation are idiosyncratic and promotional for those who complete them, a standard template for a publicly shared safety case would foster transparency and comparability.

Recommendations

AV developers and the larger AV research community should continue to advance and integrate leading measures, including roadmanship. An industry consortium could help, including by becoming a champion for advancing roadmanship measures.

Developers should forge uniform and transparent approaches to presenting evidence for meeting safety thresholds. There are differences across the sets of circumstances in which AVs are developed (the so-called *operational design domains*, or ODDs, meaning the environment in which the AV is designed to operate automatically), use cases, business models, technology, and so on. One aspect of establishing uniformity could be agreement on a safe or better-than-average driver as a threshold for commercial deployment, with concomitant efforts on behalf of AV developers and government to define conceptually and quantitatively what this threshold means within each AV's ODD.

The Department of Transportation should support further research into human drivers to enable the comparisons that stakeholders seek. This support could include collaboration with other federal agencies and private companies. Examples include understanding how people drive in different ODDs; near misses by ODD; and how to characterize a safe, rather than an average, human driver.

AV developers should collaborate with state and local leaders to bring their vehicles into communities around the country. As interviewees noted repeatedly, seeing is believing. Gaining more visibility for AVs is an important stage in building public confidence and promoting adoption. Collaborations could include demonstrations of how AVs operate, things AVs can do that might improve on human perception and reaction capabilities, and the nature and implications of ODDs, including what AV passengers and other road users need to know.

Acknowledgments

Our research team is grateful to the many individuals who shared their time and perspectives for this project, both the anonymous members of the American Life Panel who responded to a novel survey and the broad array of people with expertise relating to automated vehicles who agreed to be interviewed by the team. Those experts included executives or officials from the Advocates for Highway and Automotive Safety, the American Association of State Highway and Transportation Organizations (AASHTO), the Arizona Department of Transportation, the Arizona Institute for Automated Mobility, Aurora, Autonocast, the California Department of Transportation, Consumer Reports, Edge Case Research, the Insurance Institute for Highway Safety (IIHS), Intel, KPMG, Metamoto, Motional, Motus Ventures, the National Institute of Standards and Technology (NIST), the National Safety Council (NSC), the National Transportation Safety Board (NTSB), NVIDIA, the Pennsylvania Department of Transportation, the San Francisco Metropolitan Transportation Authority, SAE International, the Toyota Research Institute, the Uber Advanced Technologies Group, Voyage, Waymo, and Zoox, as well as transportation expert Jane Lappin, legal scholar Bryant Walker Smith, and a European technologist who preferred to remain anonymous.

As part of RAND's quality assurance process, Liisa Ecola and Jane Lappin provided constructive feedback that helped us sharpen our analysis and exposition. Within the RAND Corporation, a variety of colleagues provided essential support, including James Anderson and Nidhi Kalra, who pilot-tested the interview protocol; David Adamson and Arwen Bicknell, who helped to enhance the flow of the text; and the American Life Panel team who implemented the survey of the general public.

Our team could not have undertaken this project without the financial support of the Uber Advanced Technologies Group, and we appreciate the encouragement provided by Chan Lieu, along with Nat Beuse and colleagues, over the course of our work. We are responsible for the work presented in this report.

Abbreviations

ADAS	Advanced Driver Assistance Systems
ADS	automated driving system
ALARA	as low as reasonably achievable
ALARP	as low as reasonably practicable
ALP	American Life Panel
AV	automated vehicle
AVSC	Automated Vehicle Safety Consortium
COVID-19	coronavirus disease 2019
FDA	U.S. Food and Drug Administration
FMVSS	Federal Motor Vehicle Safety Standards
GAMAB	<i>Globalement Au Moins Aussi Bon</i> [French: “generally at least as good as”]
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
MEM	minimum endogenous mortality
mph	miles per hour
NHTSA	National Highway Traffic Safety Administration
ODD	operational design domain
PAS	Publicly Available Specification
PAVE	Partners for Automated Vehicle Education
RSS	Responsibility-Sensitive Safety
TAM	technology acceptance model
UTAUT	Unified Theory of Acceptance and Use of Technology
VMT	vehicle miles traveled

1. Introduction

Automated vehicles (AVs) promise a safer driving experience than do conventional, human-driven vehicles. Safety has informed the development and selection of specific AV features, the automated driving system (ADS) overall, and subsequent improvements to the ADS. Yet assessing AV safety has proven to be a complex undertaking. Establishing the baseline level of AV safety and the improvement required for AV safety over time are particularly complicated. People cannot “look under the hood” to understand the basis for any given developer’s claims about safety.³ Meanwhile, experiences across the industry have led developers and analysts to extend time lines for commercial deployment, signaling recognition of greater difficulty than first understood.

Gauging AV safety presents challenges. Many stakeholders who might welcome the benefits of AVs have concerns about how to define AV safety and seek to avoid the costs of insufficient safety. Those stakeholders include AV developers (both new, tech-based companies and companies with a long history of conventional vehicle production), industry analysts, safety and consumer advocacy organizations, government officials (federal, state, and local), safety researchers, and the general public (people expected to ride in or share the road with AVs). This report documents our quest to understand what might be involved in stakeholders deeming AVs acceptably safe. Our methods, including how we collected input from stakeholders, appear at the end of this chapter.

Defining Safety Is Problematic

The concept of safety is contextually, technologically, and culturally dependent. Such differences affect the details of AV design and engineering, but not, as one technologist interviewed for the project explained, the fundamentals:

Ask multiple engineers about safety, and they may give you different answers, but they will be related. Safety is different in every environment, so one first must be clear about the definitions of safety. Safety may be different for every manufacturer, but companies must define what is safety for them.

Because there is no single definition for safety in general or in transportation, this report, like its predecessor, *Measuring Automated Vehicle Safety: Forging a Framework* (summarized in Box 1.1),⁴ uses a working definition of *safety* as avoiding harm to people, whether they are in AVs or near them.

³ As one person consulted for this project observed, “When a human has a driving test, the evaluator is inside the vehicle. You can’t do that with AVs.”

⁴ Laura Fraade-Blanar, Marjory S. Blumenthal, James M. Anderson, and Nidhi Kalra, *Measuring Automated Vehicle Safety: Forging a Framework*, Santa Monica, Calif.: RAND Corporation, RR-2662, 2018.

Box 1.1. Summary of *Measuring Automated Vehicle Safety: Forging a Framework*

Measuring Automated Vehicle Safety: Forging a Framework,^a the predecessor to this report, addresses how to define safety for AVs, how to measure safety for AVs, and how to communicate what is learned or understood about the safety of AVs. Given how limited AVs' total on-road exposure is compared with that of conventional, human-driven vehicles, that report discusses options for proxy measurements—i.e., factors that might be correlated with safety—and explores how safety measurements could be made.

The report also presents a structured way of thinking about how to measure safety in a technology- and company-neutral way, and it proposes a new kind of measurement. The methods of measuring safety must be valid, feasible, reliable, and non-manipulatable. They can be leading measures (i.e., proxy measures of driving behaviors correlated to safety outcomes) or lagging measures (i.e., actual safety outcomes involving harm). The report describes a new kind of leading measure, called *roadmanship* (i.e., the ability to drive on the road safely without creating hazards and responding well to hazards created by others) that captures how well an AV behaves in traffic.

As suggested by its title, *Measuring Automated Vehicle Safety: Forging a Framework* describes a framework that shows measurement possible in different settings (simulation, closed courses, and public roads with and without a human safety driver) and at different stages (development, demonstration, and deployment). While acknowledging that the closely held nature of AV data limits the amount of information that is made public or shared between companies and with the government, the report highlights the kinds of material that could be presented in consistent ways in support of public understanding of AV safety. Clearer communication between the industry and the public about safety will be critical for public acceptance of AVs. The more consistent the communication about AV safety from industry, the more cohesive and comprehensible the message will be. *Measuring Automated Vehicle Safety: Forging a Framework* provides defined concepts (particularly those relating to safety measurement), insight into AV development and testing processes, and observations about the difficulties of assessing AV safety that inform this report.

In addition to the framework, other recommendations from *Measuring Automated Vehicle Safety: Forging a Framework* include the following:

- Pursue the opportunity to leverage a demonstration stage as a time for communication outside a company about safety (e.g., to policymakers or the public).
- Treat safety events that arise before the accumulation of exposure sufficient for statistically meaningful comparisons as case studies that can contribute to broad learning across the industry and by policymakers and the public.
- Given the potential for broader learning across industry and government, encourage development of a protocol for information-sharing.
- Create a taxonomy for common use that facilitates understanding of and communication about operational design domains (ODDs).
- Conduct more research on how to measure and communicate AV system safety in an environment wherein the system evolves through frequent updates.

^a Fraade-Blanar et al., 2018.

Safety as a Compound Concept

Safety is often addressed as a quality shaped by hazards, risks, consequences, severity, and uncertainty.⁵ All of these features relate to understanding how likely it is that something bad might happen and how bad that something might be. The global grappling with the coronavirus disease 2019 (COVID-19) pandemic as this report was written underscores questions about baseline and novel risks in people's lives. Some of these are out of an individual's hands; others

⁵ For an exploration of the concepts and their complementarity, see Niklas Möller, Sven Ove Hansson, and Martin Peterson, "Safety Is More Than the Antonym of Risk," *Journal of Applied Philosophy*, Vol. 23, No. 4, 2006.

can be modulated by an individual's behavior. As one person mused to the research team before everyday activities were constrained because of the pandemic:

What kinds of risk do we accept as we walk out the door, go down the street, get on a city bus, go into a store—what is the existing ambient level of risk that we currently accept as normal, and what change does the introduction of highly automated vehicles create? Can we characterize the kinds of change it creates, put a boundary around it?

People willingly take risks, and (as discussed later in this report) they value safety in different ways. As people make decisions about whether a new technology is acceptable (to themselves as individuals or to society), they are evaluating the technology for how risky they perceive it to be, how much trust they have in the source of the technology, and the benefit they believe they will receive from the new technology. Each of these evaluations is modified by the amount of knowledge a person has about that technology and other individual differences, including in overall attitude toward emerging technologies.⁶ Perceptions of risk can be at odds with the evidence.

Safety Is a Spectrum

Today, no transportation technology is completely safe.⁷ Even technologies designed to add safety to transportation, such as airbags, have been associated with a low level of risk; an apt comparison would be to food additives that are “generally recognized as safe” (a term enshrined in regulation) but also have side effects.⁸

Various methods can be used to understand when and how new technology might be safer than what it is replacing, as well as when and how new risks might emerge. In both cases, the process is imperfect and subject to estimation and assumptions. Quantitative and qualitative approaches can be used to understand the level of safety in a given context, although assertions about safety involve some degree of estimation and assumptions.⁹ As one person consulted for this project observed, “Safety is not black and white, and moving to objective terms from subjective, it becomes a very hard problem.”

⁶ Nidhi Gupta, Arnout R. H. Fisher, and Lynn J. Frewer, “Socio-Psychological Determinants of Public Acceptance of Technologies: A Review,” *Public Understanding of Science*, March 1, 2011.

⁷ That is captured in such adages as “the only train that never crashes is the one that never leaves the station.” Experts consulted for this project shared a few such expressions.

⁸ U.S. Food and Drug Administration (FDA), “Generally Recognized as Safe (GRAS),” webpage, September 6, 2019b.

⁹ Quantitative and formal methods are employed by engineers and statisticians and data scientists.

AV Safety in Context

AVs are not being developed in a vacuum. In the past half-century, there have been many endeavors to make human-driven vehicles safer. For conventional automobiles, the push toward safer vehicles has led the industry to adopt new technologies, such as seat belts, side airbags, and crumple zones—sometimes in response to regulatory standards and/or public pressure and sometimes voluntarily.

For example, Vision Zero is a global movement aimed at eliminating traffic fatalities. Progress toward Vision Zero’s goal includes the continued adoption and development of technologies to improve the safety of human-driven cars, many featuring some degree of automation, such as adaptive cruise control, automated antilock braking systems, and lane-keeping assistance. Vision Zero involves a systems perspective that goes beyond vehicles to include infrastructure, behavior of people on or near roadways, law and policy, and more.¹⁰ That systems perspective has a place in AV development, as one person interviewed for the project outlined:

I get asked, “what can [government officials] do to improve safety and stop the loss of loved ones?” We have seen the addition of safety equipment in cars, changes outside cars and to infrastructures. We haven’t been able to tackle the operator, the individual controlling the vehicle—we tried punishment, incentives, [but because of] health issues, substance abuse, and distractions, people are a danger to themselves or on the roadway. There was an opportunity eight or nine years ago to see what technology can do to address operation of the vehicle. If we can get to the point where . . . we are down to a . . . few [operating systems], we can tackle the issues. . . . There is a lot of promise.

Consistent with Vision Zero, this quotation illustrates the hope that AVs can improve safety outcomes compared with those of fallible humans; it also acknowledges that progress will involve maturation, including consolidation, of the AV industry.

The context for AV development centers on the industry producing AVs, which is evolving along with its product. The originally separate branches associated with the tech sector and conventional vehicles have begun to exchange people, capital, and practices, and some consolidation is occurring. The risks to public safety when a small number of vehicles is being tested in a few locations are different from the risks associated with the deployment of numerous vehicles in many places.¹¹ The risks are also different when the AV business model assumes fleet ownership and operation as opposed to personal ownership prevalent with human-driven cars.¹²

¹⁰ Liisa Ecola, Steven W. Popper, Richard Silbergliitt, and Laura Fraade-Blanar, *The Road to Zero: A Vision for Achieving Zero Roadway Deaths by 2050*, National Safety Council (NSC) and RAND Corporation, RR-2333-NSC, 2018.

¹¹ For one view of how the diffusion of AVs might arise from a collection of discrete early-adopting locations, see Gary Silberg, *Islands of Autonomy*, Amstelveen, Netherlands: KPMG, 2017.

¹² For example, ride-hailing or ride-sharing services involve fleet operation and could feature fleet ownership.

Perhaps the most important context is the environment (defined by geography, terrain, weather, lighting, traffic, roadway complexity, and more) in which the AV is tested and for which it is designed: the ODD. In simple terms, the development of an AV begins with a small and simple ODD that expands and grows more complex as ADS technology improves. Ultimately, a fully automated ADS (level 5 in the SAE scale¹³) could operate anywhere and anytime, but the ODD is generally limited, including for highly automated AVs (level 4). The ODD also reflects local driving culture (which varies within and among countries) and infrastructure (which also varies and evolves). This report focuses on level 4 and level 5 vehicles.

A last dimension of AV context to note is that the total fleet of vehicles on the road continues to evolve. The number of AVs being tested on public roads at the time of this writing is a tiny fraction of the total fleet.¹⁴ But with continued development, understanding, and authorization (by government at different levels), the expectation is for a growing percentage of the total fleet to be automated. Definitions of safety will vary with that percentage, another kind of systems problem that will engage developers, engineers, and policymakers.¹⁵

“Acceptably Safe” and This Report

Given the difficulties of assessing AV safety, the question arises, how safe is acceptably safe, or “safe enough,” for AVs? One person interviewed for the project peeled back some of the layers of this onion and illustrated how risk management contributes to judgments about whether AVs are acceptably safe:

We look at *safe enough* in the context of the situation, such as maturity and life cycle. . . . Today we are developing technology to understand which approaches work, or don’t, or are promising. What is safe enough for prototype vehicles? This is a question that is important to answer before we get to what is safe enough for deployment. When thinking about safe enough for deployment, one must consider ODDs—specific environments. Having defined what is safe enough, and how much safety that is being aimed for, then you have to figure out if what you have done actually meets those targets and how the AVs will continue to do so as they are continually developed. Without considering those specifics, you can’t say AVs are safe enough.

For instance, an automated driving system should not be the cause of an accident. That sounds nice, but an engineer can’t do anything with that definition. Further specifications, such as “don’t get hit” or “don’t endanger others,” make things

¹³ SAE International, “Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles,” Standard J3016_201806, June 15, 2018. SAE, formerly the Society of Automotive Engineers, is now known by the abbreviation alone. This standard provides a taxonomy of automated driving capabilities, ranging from level 0 (no driving automation) to level 5 (fully automated driving with no human needed).

¹⁴ There are more than 250 million registered vehicles in the United States, a fleet with a wide mix of features and safety expectations.

¹⁵ Developers expect that a fleet of AVs will be safer than a fleet combining AVs with significant numbers of human-driven vehicles, in part because the unpredictability of the human factor will have been eliminated.

more specific, but they aren't good enough, either. How do you not get hit? Hit by what? Et cetera. It's difficult to connect the high concepts of safety with the on-the-ground aspects.

Others noted a chicken-and-egg problem (“‘safe enough’ is still entirely hypothetical—it remains undefined, and will remain so until there is much wider roll-out of AVs”). Another of our interviewees argued that “‘safe enough’ was never defined for automobiles in the first place, which rolls over into the difficulty of defining it for AVs.” Observed pragmatically, “[f]rom a public acceptance perspective, ‘safe enough’ and ‘acceptably safe’ are the same. . . . Both terms present a certain amount of ‘unsafe,’ of risk.” One person argued against anything other than “safe”:

We're trying to move away from the term “safe enough” to talk about “safe.” Part of the challenge we have is defining “safe” today, because there really isn't a definition for what “safe” is today. There are lots of examples—you can point to scales or penalties when people get in accidents or points on their license, but they're still “safe enough” to drive.

Rather than debate semantics, we consider both questions in this report—what makes AVs acceptably safe and safe enough. We set out to explore both different stakeholder perspectives on what it means for an AV to be acceptably safe and the potential for agreement across stakeholders. Although stakeholders think differently about safety, the report is intended to be useful to a variety of stakeholders.

Our collection and analysis of inputs through methods outlined in the following sections focused on understanding how different stakeholders perceive, appraise, understand, compare, and/or use different kinds of evidence for assessing AV safety. We considered both the nature of the evidence and how it might be communicated.

Mixed Methods

Three principal activities shaped the research that went into this project. These three lines of effort are described in more detail in the appendixes to the report.

First, we collected and reviewed literature, including both scholarly literature and gray literature (government, and other nonscholarly publications) of different kinds. We continued to collect and review literature throughout the project, providing theoretical, empirical, and other context.

Second, we conducted a survey of public views on sources of evidence for AV safety using the RAND American Life Panel (ALP), a nationally representative sample of the U.S. population. The survey provided insight into how the general public weights different evidence types and sources relating to AV safety. Unlike most public opinion surveys about AVs, it did not poll people on their opinions about AVs or the prospect of riding in them.

Third, we conducted a series of semistructured interviews with a wide variety of stakeholders other than the general public. Although specific comments are not attributed to interviewees (except when authorized), many quotations appear throughout the report to convey the variety

and flavor of perspectives collected. These unattributed comments dominate the source material for this report, complemented by the ALP survey data and augmented by the literature.

Limitations

In a new, rapidly evolving, competitive industry that is advancing and using complex technology, there is no substitute for talking with people engaged in work in that industry. Although we benefited enormously from the many interviews we conducted, our generous sample was not a complete census of even U.S. developers. We also sampled other kinds of stakeholders. We were pleased with the distribution of interviewees, but we recognize that we could have learned more from an even larger number of interviews.

Organization

The rest of this report is organized as follows: Chapter 2 introduces the approaches that we examine in more depth in Chapters 3–5 (a structure that necessarily involves a degree of repetition). In Chapter 3, we examine safety assessment using approaches to measurement. Chapter 4 describes how processes used in industry or government can serve to assess safety, complementing or providing an alternative to measurements. Chapter 5 addresses thresholds as an approach to assessing safety, discussing methodological challenges and relating thresholds to measurements and processes. In Chapter 6, we consider the issues associated with communication about risk and safety. Chapter 7 presents conclusions and recommendations.

In addition to the appendixes on methodology and the literature, a bibliography combines references cited with other literature that we consulted and found useful for understanding the topic.

2. Safety as a Codified Concept for AVs

In this report, we examine different approaches to assessing AV safety. Transportation can never be completely free of risk, and the parameters of acceptable risk captured by different approaches can be helpful for risk management.¹⁶ Approaches need not be independent of each other, and they can be combined to yield additional information. Different approaches reflect how we as individuals or as a society assess, contextualize, and judge the safety of AVs both in their existing form and against our expectation for them to become acceptably safe.

In Table 2.1, we present different ways to categorize AV safety approaches: safety as a measurement, safety as a process, and safety as a threshold.

Table 2.1. Approaches for Assessing AV Safety

	Safety as a Measurement (Chapter 3)	Safety as a Process (Chapter 4)	Safety as a Threshold (Chapter 5)
Description	<ul style="list-style-type: none">• A quantitative measure authenticating the safety performance of AVs using data-driven evidence	<ul style="list-style-type: none">• Indicators of developer behaviors consistent with achievement of safety and sometimes connecting to quantitative measurements. This approach includes the engineering efforts involved in system-level validation and verification that, in conjunction with other practices, contribute to a safety case. A safety case might be associated with any or all of the three process approaches.	<ul style="list-style-type: none">• The AV's performance, determined through either a measurement or a process, is shown to be at a given level.
Variants	<ul style="list-style-type: none">• Safety as indicated by a leading measure (i.e., measures of pre-crash driving behavior)• Safety as indicated by a lagging measure (i.e., measures of crashes and post-crash outcomes)	<ul style="list-style-type: none">• Safety as indicated by compliance with technical standards or best practices• Safety as indicated by compliance with federal, state, or local regulations• Safety as indicated by corporate safety culture (relating to the pervasiveness of attention to safety across a developer and its activities and personnel)	<ul style="list-style-type: none">• Safety as indicated by meeting a goal based on human-driving performance (focuses on a threshold predicated on some level of human driving)• Safety as indicated by meeting a goal based on ADS technology potential. This threshold includes scenario-based driving tests.

¹⁶ An illustration for water safety, which uses the term “standpoints” in discussing different approaches, can be found in Lorna Fewtrell and Jamie Bartram, eds., *Water Quality: Guidelines, Standards and Health*, London: IWA, 2001.

None of these approaches stands alone. Rather, they build on each other. For example, an AV reaching a given safety threshold (from the safety as a threshold approach) can be supported by evidence from safety measures and processes showing progress toward that threshold. Achieving AV safety is a not one-time event; rather, it is an ongoing process, and evidence of safety provides a staircase to achieving each new threshold. (Figure 2.1).

Figure 2.1. Approach Relationships



Any approach for AV safety must be valid and feasible. The former means that the approach must actually reflect safety (rather than, for example, driving function, public relations, or technology development). The latter means that it must be possible to generate evidence to support (or oppose) meeting the approach with reasonable effort and resources—there has to be a practical implementation, not just a theoretical construct. Generating an approach might require additional testing, data collection, or processes beyond what would normally be performed.

Safety as a measurement focuses on new measurements or others that are not feasible for conventional vehicles. The *processes* we examine are either those specifically created for AVs or those previously applied to conventional vehicles. The *thresholds* are either based on human-driving performance or unrelated to human drivers. The approaches here are unique because AVs are new to the transportation ecosystem and not a simple evolution of in-vehicle technology.

We do not anticipate that every company will use the same approaches to assess and articulate product safety, nor that companies will use the same technique within each approach.

Rather, evidence of an acceptable level of safety is specific to the ADS, the ODD, the business model, and the company's capabilities and resources. Although approach definitions are uniform, methods for applying them might not be. This presents complications. Not all methods are equally convincing. For example, crash evidence from one suburban block for 1,000 hours might be more or less convincing than evidence from driving a simulated suburban block for 5,000 hours or evidence from driving an entire suburban community for 500 hours. We discuss these challenges in the following chapters.

Regardless of which approaches a given company uses, they must be communicated effectively to regulators, safety advocates, the general public, and others (Chapter 6). Statements conveying information and/or endorsements from various sources (AV companies, government, safety advocates, and others) aim to explain a new type of transportation technology to the public. Statements provide information about the capabilities of AV technology and about appropriate usage and expectations of said technology—and, ultimately, these statements pave the way for consumer engagement.

3. Safety as a Measurement

Measurements are established forms of evidence for determining acceptable safety. For example, event rates per *exposure* (such as vehicle miles traveled [VMT] or hours driven) have been used to show and compare safety performance. As discussed in Chapter 1, characterizing levels of risk and safety for AVs in ways that are acceptable to a wide variety of stakeholders involves (1) identifying what certain measurements can (or cannot) show about AV safety and (2) communicating this knowledge. This chapter focuses on the former, describing the knowledge that safety measurements can yield.

How Measures Show Safety: Lagging and Leading Measures

Our 2018 report, *Measuring Automated Vehicle Safety: Forging a Framework*, articulated two streams of measurement: leading and lagging measures.¹⁷ Interviews conducted for this project indicated that this categorization has been taking hold, as illustrated by the following quotation:

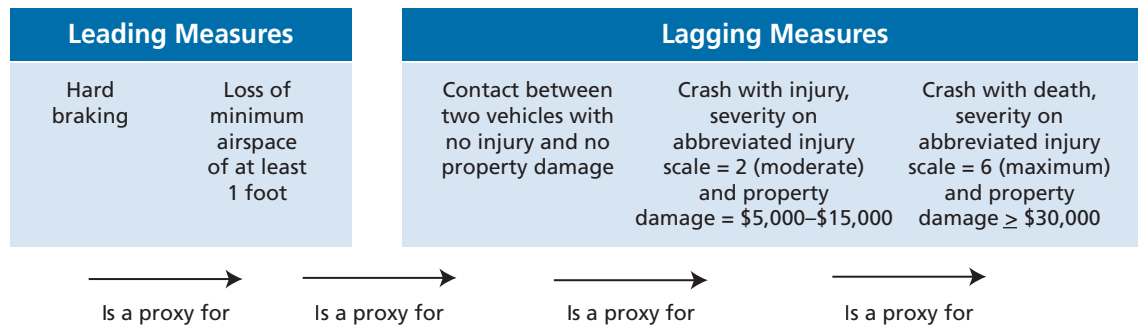
We should be able to develop specific measures relevant to safety. We've latched onto the idea of leading and lagging metrics. Leading includes general performance characteristics, associated with vehicle kinematics. Post-deployment, [we want to] measure lagging, longer-term metrics for more-societal concerns (crashes, near misses, things yet to be defined).

The 2018 report also described what it called roadmanship as a novel, integrative, leading measure.

Leading measures are pre-crash measures of prevention and of behavior; they are proxies for lagging measures. *Lagging measures* reflect crashes and post-crash outcomes, such as crashes, injuries, costs, and deaths. More broadly, driving events exist on a continuum of severity and temporality (e.g., hard braking is a proxy for a near miss of 0.5 inches, which is a proxy for a low-severity crash, which is a proxy for a medium-severity crash with injury, which is a proxy for a high-severity crash with deaths [see Figure 3.1]). The term *proxy* here is used in the statistical sense, in which information on one event (in this case, an event occurring more frequently) serves as an indicator for the likelihood of a rarer and more severe event or outcome. Thus, a lagging measure might serve as a leading measure or as a proxy of the likelihood for a lagging measure happening later in the crash sequence (the events that occur before, during, and after a crash) or with greater severity. Each measure is one piece of evidence, but none stands alone; a varied collection is necessary to get a full picture of safety.

¹⁷ Fraade-Blanar et al., 2018.

Figure 3.1. Illustrating Selected Driving Events as a Continuum



Safety as Demonstrated by Lagging Measures

Using lagging measures to characterize levels of safety is broadly understandable—especially to the general public and many policymakers. As one interviewee stated:

We need both [kinds of measure]. Leading measures are things that we would look closely at. Lagging measures are things that we would look at and say “what has been done to change this number” and assess if what happened in the past will continue to happen in the future.

Another said:

On the lagging side, what you’d like to see is a reduction of injurious and fatal crashes, and that will tell the final story on if it is safe or not.

Deaths from crashes are perhaps the best understood lagging measure, with absolute numbers reported each year as evidence of the national motor vehicle safety level. Deaths are also the best documented and easiest to count because the definition is precise and uniformly applied, and numbers are published. Nonfatal injuries and crashes are equally reflective of safety levels, but data are less available.¹⁸ Because severity thresholds for reporting vary by state, crashes, although an indicator of safety, are slightly more ambiguous. This is particularly true for low-severity crashes with minimal property damage and no injuries that are not reported to the state.

Safety as Demonstrated by Leading Measures

Lagging measures are assessed using events that are relatively rare for AVs. As earlier RAND research has noted, it would take hundreds of millions—possibly billions—of VMT to

¹⁸ Fatal crashes are documented in the Fatal Analysis Reporting System (FARS), a national census of all deaths occurring within 30 days of a motor vehicle crash as a result of said crash. Crash Investigation Sampling System (CISS)—formerly National Automotive Sampling System Crashworthiness Data System (NASS CDS)—is a sample of crashes nationwide in which at least one vehicle was towed.

generate statistically valid event rates of crashes, injuries, and deaths.¹⁹ Leading measures, which use events that occur more frequently, are the alternative.²⁰

The chief challenge with leading measures for both human drivers and AVs is that there is no established metric. For example, the shortcomings of using counts of disengagements,²¹ an AV-specific leading measure, are immense and generally agreed on. Because they are deeply rooted in circumstances (of the environment, the driver’s risk tolerance, company policy, reporting parameters, system manipulability issues, etc.) disengagements do not reflect safety. Interviewee comments were largely negative: “disengagements are useless,” “heavy discounting of disengagement reports,” “limitations of early metrics [mean that] disengagements are not a measure of safe enough—measures of whether [the ADS is] mature,” “disengagement rates will be difficult to convey to the public with a tweet or a sound bite; it’s too complicated,” “disengagements are highly gameable,” “lots of hate on disengagement reports right now.” Here are two longer illustrations from project interviews:

For years, we’ve seen people cite the California disengagement reports as the ultimate level of safety. That has been very sad, as (a) it’s just a subset in California, and (b) there is a huge gap in terms of being able to provide the public with critical analysis about how something is safe.

The problem with disengagements: There isn’t enough context around them. You can’t be sure that a disengagement for [company A] is the same as for [company B], or even for [company A] from one year to the next. Not just a hard brake, but a hard brake in response to what?

Legal infractions as a leading measure elicited little enthusiasm and were mentioned only in one interview, possibly because sometimes an illegal maneuver is the safest course and because the correlation between infractions and collisions is established to be statistically significant but not extremely strong or well-understood. With no other strong leading-measure candidates, conversation and usage around leading measures have revolved around the concept of roadmanship (see Box 1.1; the concept is also discussed in more detail in the next subsection). The interviews conducted for this project suggest that leading measures and roadmanship measures have become effectively synonymous.

¹⁹ Nidhi Kalra and Susan M. Paddock, *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* Santa Monica, Calif.: RAND Corporation, RR-1478-RC, 2016.

²⁰ Amitai Y. Bin-Nun, Anthony Panasci, and Radboud J. Duintjer Tebbens, *Heinrich’s Triangle, Heavy-Tailed Distributions, and Autonomous Vehicle Safety*, presented at the 99th Transportation Research Board annual meeting, Washington, D.C., January 2020.

²¹ A disengagement occurs when a human safety driver assumes control of the vehicle being tested, disengaging the ADS. The ADS or the safety driver can initiate a disengagement.

Roadmanship

Roadmanship as a set of leading measures was widely cited by interviewees as a promising approach to measuring acceptable safety. *Roadmanship* is an umbrella term for leading safety measures focused on safe driving behavior. Roadmanship measures whether the AV is acting according to rules of the road, executing, in the words of interviewees, “reasonable behavior on the roadway,” and/or “contributing to the harmonious flow of traffic, not impeding traffic.” Roadmanship measures reflect the ADS’s situational awareness; time to response; speed adjustment conditional on prevailing factors; and the developer’s knowledge of the vehicle’s physics, local driving culture, and local driving laws, all combined with a deep reservoir of observed and actual experience with diverse driving scenarios. Roadmanship also encompasses the trust and comfort of the passenger and other road users.

Roadmanship presents opportunities but also challenges. It “moves from ‘don’t crash’ [a lagging metric] to ‘drive safely.’ ‘Don’t crash’ is objective, ‘drive safely’ is . . . much harder to find a formal definition for.” Although the concept continues to evolve, roadmanship also continues to hold an important position as an aspirational leading measure for AV safety.

The original articulation of roadmanship focused on safe roadway citizenship. During the research for this project, stakeholder comments deepened the discussion of roadmanship:

When we’re talking about roadmanship, we’re talking about behavioral planning of the paths AVs take. . . . The roadmanship concept is much broader than just accidents. It goes into the qualitative nature of driving that allows you to avoid accidents in the first place, such as turn radius, smoothness of acceleration and turning, integrating the human norms of driving in the region, and more. . . . Roadmanship isn’t necessarily a new term, but it has many different aspects, such as the [original equipment manufacturer] side of things. The general problem is, how can we generate good and safe behavior? But there isn’t a common definition of what that means.

Another interviewee stated,

On a 1–10 scale where 10 is Vision Zero [no fatalities], roadmanship is at the 3–5 level. How do we get into the second half of the continuum to get prevention down? We need to figure out what more to do. Some things are appearing that are good, like RSS [Responsibility-Sensitive Safety]. The top part of the continuum is the prevention stuff, proactive part. Avoiding situations that could potentially kill anybody. People looking at level 4 and level 5 vehicles are saying that the trolley problem²² [involving ethical quandaries] should never be an issue—the ADS will avoid the situation.

²² The trolley problem is a classic philosophical thought experiment: One is on a trolley with a broken brake. If it continues on its current track, it will run down and kill a group of people with certain characteristics. If one throws a switch, the trolley will be diverted onto another track, killing another group of people with certain characteristics. What does one do? (Aarian Marshall, “What Can the Trolley Problem Teach Self-Driving Car Engineers?” *Wired*, October 24, 2018; Lauren Cassani Davis, “Would You Pull the Trolley Switch? Does It Matter? The Lifespan of a Thought Experiment,” *The Atlantic*, October 2015)

Central to roadmanship measures is the ability to signal safety by a tight correlation with lagging measures. Table 3.1 illustrates the relationship between one possible measure of roadmanship—hard braking—and crashes. Each cell tells a different story about roadmanship’s ability to serve as an approach to safety. The goal is for hard braking to occur when danger is present (cell A) and no hard braking to occur when danger is absent (cell D). Cell D is the optimal situation, in which the AV either avoids a potentially dangerous situation or is able to handle all danger without taking evasive maneuvers. Cell A is both a success and a failure: It is a success because the system reacted appropriately and prevented a crash and a failure because the ADS failed to avoid a potentially dangerous situation, resulting in a leading-measure event occurring.

Table 3.1. Relationship Between Hard Braking and Appropriate ADS Action

Braking Event	Danger Present	Danger Absent
Hard braking occurred	A: appropriate reaction	B: false positive
No hard braking	C: false negative (crash)	D: appropriate avoidance

A false positive (cell B) is uncomfortable for the passenger but not inherently dangerous. A false negative (cell C, no hard braking when danger was present), likely means that either (1) the danger presented too quickly for the system to react (the crash was not preventable) or (2) the ADS did not recognize the danger in time to react (the crash was preventable).

Part of the challenge of measuring roadmanship is that there is no consensus definition for good driving behavior. As one interviewee said, “[I] know what it is but I can’t say what it is.” Another stated:

Roadmanship means “good behavior,” but we still don’t know what that means. We can find a definition to talk about the concept, but what are the specifics? There are a lot of arguments about if it is possible and if it should be pursued. It is very different between Tel Aviv, Delhi, and Detroit. We tend to agree on the physics parts, but the road rules and behaviors are one of the big things . . . [where agreement is needed]. What are the points where we can realistically make assumptions?

There are ongoing efforts from such groups as Aptiv’s Structured AI [Artificial Intelligence] system to generate foundational and mathematically defined driving rules an AV could follow,²³ something that the algorithms of RSS (discussed below) can also provide.²⁴ Although these are

²³ Andrea Censi, Konstantin Slutsky, Tichakorn Wongpiromsarn, Dmitry Yershov, Scott Pendleton, James Fu, and Emilio Frazzoli, *Liability, Ethics, and Culture-Aware Behavior Specification Using Rulebooks*, Dublin, Ireland: Aptiv, white paper, 2019.

²⁴ Shai Shalev-Schwartz, Shaked Shammah, and Amnon Shashua, *On a Formal Model of Safe and Scalable Self-Driving Cars*, arxiv.org, 2017.

useful, development is still ongoing, and use (and agreement that these are the right rules) is neither widespread nor uniform.

Concomitantly, there is no measurement that encapsulates all the qualities that a measure of roadmanship should contain: physics-based, objective, available using current technology, and able to distinguish the initiator from the responder, so as not to penalize a road user for taking evasive maneuvers.²⁵

In the absence of a perfect measure of roadmanship, leading measures tend to fall into a few nonexhaustive categories. Table 3.2 presents a selection of such measures. These measures could arise during scenarios run in simulation, in closed-course driving, or during on-road driving. Table 3.2 illustrates the challenge for roadmanship: No single measure covers all aspects of roadmanship, nor does any measure cover all possible threats to safety that occur on the roadway. No one measure can be used in isolation; each measure's contribution is best voiced in harmony. As one interviewee remarked:

Some of the conversations that are happening now, including around leading metrics (violation of safety envelope, etc.), no matter how they decide to measure it, we count it as a metric. We can combine that with Delta-V, TTC [time to collision], and measures of jerk to come up with a scale or rating to get a better idea of the future safe performance of that vehicle in its ODD.²⁶

Research is needed to understand how to best integrate measures of roadmanship so that the weaknesses of one component measurement are offset by the strengths of another to create a comprehensive picture of safety.

Measuring Automated Vehicle Safety: Forging a Framework notes the potential of violations of the safety envelope as a measure of roadmanship.²⁷ Although strong, this approach on its own is not sufficient to reflect roadmanship. Roadmanship has to include the vehicle's integration into existing traffic without causing any untoward disruption to established traffic flow. Unpredicted disruptions degrade safety for all surrounding road users by requiring other drivers to accommodate or to take evasive actions. As one interviewee observed:

If you pull out in front of someone, and they have to hit the brakes, you were not a nice driver. [There might be] no accident, they may have stopped before the [edge of the] "safety envelope," but you still impeded driving.

²⁵ This quality of roadmanship draws distantly from but is not the same as legal liability.

²⁶ Again, *safety envelope* refers to a desired separation of the vehicle from other objects factoring in the travel path. *Delta-V* refers to change in velocity, and *jerk* refers to the quality of change in position of the vehicle. These and other methods familiar to engineers are typically expressed and measured mathematically.

²⁷ Fraade-Blanar et al., 2018.

Such circumstances are plausible and reflect a concern germane to roadmanship, and yet it is unclear whether any of the measures in Table 3.2 would capture it correctly.²⁸

In the meantime, the availability of systems for gauging safety envelope violations has grown. Among other important developments, RSS, discussed in the previous report, has grown in sophistication; Safety Force Field has emerged; and the Instantaneous Safety Metric has continued to evolve.²⁹ In particular, RSS and Safety Force Field are also being discussed as safety checkers. Checkers are collision avoidance and collision checking systems that run parallel to the primary software.³⁰ Checkers use quantitative descriptions of safety to know when safety is under threat, “helping define what safe driving means.”³¹ Measuring safety by counting incursions into the safety envelope (or through some other means) can generate rate-based measures, such as counts per intervention per VMT. It could also contribute to implementing other measures articulated in Table 3.2.³²

Additionally, roadmanship measures are conceived of and measured at the vehicle level for the ADS, but AVs are a new component of a complex transportation ecosystem. As a tiny minority of road users, AVs are unlikely to affect the flow of traffic. But as AVs increase in proportion of the total fleet, measuring safety at the transportation system level becomes important. Potential measures include throughput of traffic used in association with overall crash rates per road users (to assess not just changes in safety but also volume-based efficiency of the overall transportation system).

²⁸ One of the involved vehicles in this example would register a hard braking event, but because that event occurred in response to another vehicle’s poor behavior, recording it would unfairly penalize the driver taking evasive maneuvers to avoid a crash.

²⁹ Frank Barickman, Joshua L. Every, Bowen Weng, Scott Schnelle, and Sugghosh Rao, “Instantaneous Safety Metric,” NHTSA, PowerPoint presentation, June 25, 2019; Joshua L. Every, Frank Barickman, John Martin, Sugghosh Rao, Scott Schnelle, and Bowen Weng, “A Novel Method to Evaluate the Safety of Highly Automated Vehicles,” presentation at the 25th International Technical Conference on the Enhanced Safety of Vehicles (ESV), National Highway Traffic Safety Administration (NHTSA), Detroit, Mich., 2017; Bowen Weng, Sugghosh Rao, Eeshan Deosthale, Scott Schnelle, and Frank Barickman, *Model Predictive Instantaneous Safety Metric for Evaluation of Automated Driving Systems*, arxiv.org, May 2020.

³⁰ A *safety checker* is a system that adds redundancy and complements the primary system (which is the *doer* in the doer-checker dyad) and in effect constantly asks whether the doer is, as the team heard in interviews, “always making the right decisions” and “checking for imminent collision and take action if something failed in the primary path.” These checkers provide an additional layer of protection against mistakes by the artificial intelligence, which is “probabilistic not deterministic.”

³¹ Junko Yoshida, “Can Mobileye Validate ‘True Redundancy’?” *EE Times*, May 22, 2018; Philip Koopman, Beth Osyk, and Jack Weast, *Autonomous Vehicles Meet the Physical World: RSS, Variability, Uncertainty, and Proving Safety*, arxiv.org, 2019.

³² Jeffrey Wishart, Steven Como, Maria Elli, Brendan Russo, Jack Weast, Niraj Altekar, and Emmanuel James, *Driving Safety Performance Assessments Metrics for ADS-Equipped Vehicles*, SAE International, Technical Paper 2020-01-1206, 2020.

Table 3.2. Strengths and Weaknesses of Selected Measures' Abilities to Reflect or Fail to Provide Evidence of Acceptable Safety

Characteristic	Measures of Behavior		Measures of Time	Measures of Safeguard Engagement	Measures of Probability	Measures of User Perception
Example	Rapid acceleration or deceleration ^a	Safety critical events ^b (combination of crashes and near crashes)	Time to collisions ^c (time to crash, if vehicles continue on current paths)	Safety envelope violations ^d (airspace or other violations resulting in engagement of safety checkers; e.g., RSS or Safety Force Field)	Instantaneous safety metrics ^e (probability of an unavoidable crash, recalculated at every instant)	User-identified near misses ^f (an incident involving a subjective judgment of crash likelihood)
Strengths						
Easily communicated	X	X	X		X	X
Already in use	X	X	X	X		X
Objective	X		X	X	X	
Uses statistical assumptions that can be validated	X		X	X	X	
Weaknesses						
Lacks a universal definition	X			X		X
Doesn't cover all crash types	X					
Can create perverse incentives	X					X
Time-intensive to generate		X	X		X	X
Uncertain correlation to crash		X ^g				X
Uses statistical assumptions that can be faulty	X		X	X	X	
Might need infrastructure investment			X			
A work in progress				X	X	

NOTES: This is not an exhaustive list; many of these categories have other measurements included. For example, measures of time also includes post-encroachment time. For more information, please consult Fraade-Blanar et al., 2018. This table represents selected measures of roadmanship. Inclusion was based on interviewees presenting the measures they saw as having the widest industry consideration and/or the greatest potential. For a more comprehensive list of safety measures, see Wishart et al., 2020.

^a Yuji Arai, Tetsuya Nishimoto, Yukihiro Ezaka, and Kenichi Yoshimoto, "Accidents and Near-Misses Analysis by Using Video Drive-Recorders in a Fleet Test," in *Proceedings of the 17th International Technical Conference on the Enhanced Safety of Vehicles (ESV) Conference*, Amsterdam, 2001; Joshua Stipancic, Luis Miranda-Moreno, and Nicolas Saunier, "Vehicle Manoeuvres as Surrogate Safety Measures: Extracting Data from the GPS-Enabled Smartphones of Regular Drivers," *Accident Analysis & Prevention*, Vol. 115, June 2018; Carl Johnsson, Aliaksei Lareshyn, and Tim De Ceunynck, "In Search of Surrogate Safety Indicators for Vulnerable Road Users: A Review of Surrogate Safety Indicators," *Transportation Reviews*, Vol. 38, No. 5, 2018;

S. M. Sohel Mahmud, Luis Ferreira, Shamsul Hoque, and Ahmad Tavassoli, "Application of Proximal Surrogate Indicators for Safety Evaluation: A Review of Recent Developments and Research Needs," *IATSS Research*, Vol. 41, No. 4, December 2017.

^b This term is used by Virginia Tech Transportation Institute's Strategic Highway Research Program 2, a naturalistic data set. In the dataset, these events "(i.e., crash, near crash, crash-relevant, non- conflict, subject conflict) were manually validated and coded by trained data reductionists." Johan Engström, Andrew Miller, Wenyan Huang, Susan Soccolich, Sahar Ghanipour Machiani, Arash Jahangiri, Felix Dreger, and Joost de Winter, *Behavior-Based Predictive Safety Analytics—Pilot Study*, Blacksburg, Va.: Virginia Tech Transportation Institute, April 2019.

^c Johnsson, Lareshyn, and De Ceunynck, 2018; Mahmud et al., 2017; Federal Highway Administration, *Surrogate Safety Assessment Model and Validation: Final Report*, McLean, Va.: U.S. Department of Transportation, February 2008.

^d *Safety envelope* refers to a desired separation of the vehicle from other objects factoring in the travel path. Koopman, Osyk, and Weast, 2019; Mobileye, "Responsibility-Sensitive Safety," webpage, undated; NVIDIA, "Planning a Safer Path," webpage, undated.

^e Weng et al., 2020; Every et al., 2017.

^f John C. Hayward, "Near Miss Determination Through Use of a Scale of Danger," in *Proceedings of the 51st Annual Meeting of the Highway Research Board*, Washington, D.C., 1972; NHTSA, *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*, Washington, D.C.: U.S. Department of Transportation, DOT HS 810 594, 2006; Sheila G. Klauer, Feng Guo, Bruce G. Simons-Morton, Marie Claude Ouimet, Susan E. Lee, and Thomas A. Dingus, "Distracted Driving and Risk of Road Crashes Among Novice and Experienced Drivers," *New England Journal of Medicine*, Vol. 370, No. 1, 2014; Nobuyuki Uchida, Maki Kawakoshi, Takashi Tagawa, and Tsutomu Mochida, "An Investigation of Factors Contributing to Major Crash Types in Japan Based on Naturalistic Driving Data," *IATSS Research*, Vol. 34, No. 1, 2010; Arai et al., 2001.

How Lagging and Leading Measures Work Together to Provide Evidence of Safety

One piece of evidence alone cannot provide proof of acceptable safety. Ideally, both leading and lagging measures are collected in three settings (simulation, closed courses, and public roads) as described in the predecessor report.³³ The utility of leading measures lies in their ability to help predict lagging measures. The strength of evidence of the correlation between roadmanship measures and crashes for human drivers varies based on the roadmanship measure under discussion.³⁴ Publicly available evidence of the correlation for AVs varies as well. In the words of some of the people interviewed for the project:

In the end, you need to use both [kinds of measure]. Given that leading measures are proxies for what you really want to measure, you've always got to check that, even though you are meeting the leading measures' goals, you are also meeting the lagging measures' [goals]. Leading measures don't always predict the lagging ones.

It may just be luck that a crash didn't occur. It may be that those leading measures aren't associated strongly enough with possible crashes. Leading measures do play a role, but we need to do more research to find out what is the most correlated with the lagging measures. This is hard to do.

AVs have different behaviors and different failure modes than human drivers.

Do you design the AV to operate like humans that constantly break one-third of driving rules, or do you operate strictly along the rules? Law enforcement will get complaints here that AVs stop at stop signs and go the speed limit, causing hazards on the roadway. It is hard when we write in our law books that you have to do these things, but then people complain that people are driving slow, meaning slower than me, slower than 90 mph [miles per hour], in the left lane.

³³ Fraade-Blanar et al., 2018. Leading and lagging measures can be gathered from a range of settings, including simulation, closed-course test tracks, and public roads (the real world), as discussed in the predecessor report. How measures reflect safety might differ within a setting. For example, in real-world driving, crash metrics might be more reflective of ADS safety performance when there is no safety driver present to disengage. Of course, overall mileage driven does not reflect the safety or maturity of the ADS. Although testing in each setting serves to complement and refine the accuracy of testing in the others (e.g., closed-course testing can be used to validate simulation) and each can be checked for every major system update, a blend is used to show evidence of safety. Furthermore, qualification of evidence based on setting is necessary—the most-valid evidence derives from real-world driving, in which the vehicle proceeds within all possible dimensions of the stated ODD and surprises are most likely. As one person interviewed for the project put it, there are “so many edge cases and situations that even experts cannot anticipate. [We] need testing from the real world to deal with surprises.”

³⁴ Fraade-Blanar et al., 2018.

As reflected in these observations, a case can be made that because AVs drive and make mistakes differently from human drivers, and new crash types emerge, crash proxies for human drivers might not translate into crash proxies for AVs.

According to this argument, for example, AVs can identify risks and respond faster than a human can, so a closer following distance or smaller airspace (safety envelope) around the AV as compared with what is needed for a human-driven vehicle can be safe and not a precursor to a crash. Such an argument has face validity. But even if less airspace is needed for an AV, some is still required, and even if an AV can identify and respond to threats faster than a human can, the laws of physics still require a minimum braking distance.

Hence, although the quantitative relationship between leading and lagging measures might differ for AVs from what they are for human drivers, the theoretical underpinnings hold. It follows that the results of leading and lagging measures should agree—or, as one interviewee put it, “it’s reasonable that fewer fatalities would also mean fewer injuries.” According to another:

Even though you’re meeting leading-measure goals, you need to make sure you are meeting what you wanted from the lagging metrics. It’s kind of like, if you create a new crash test, you make predictions about that. But over time, you want to know about the actual life-improving impact of the test.

There is an interpretational challenge if leading and lagging measures do not agree. The most obvious example is if the technology shows poor performance according to leading measures but positive performance according to lagging measures. The explanation could be simply that not enough miles have been accumulated to show poor lagging measures for AVs. The outcome also might reflect a highly cautious ADS or one driving in an ODD requiring a great deal of hard braking (e.g., downtown Manhattan during rush hour, when one might have to drive aggressively). It is also possible that the leading measures in question are poor proxies, an explanation that would also be true if the situation were reversed, with good leading-measure performance but poor lagging-measure results. Another kind of challenge arises if AVs have a lower fatal crash rate than human drivers have but a higher injury crash rate or a higher fatal crash rate for only one specific road user group (e.g., tandem-bicycle riders). Such a situation is arguably safer but not acceptably safe.

Safety I and Safety II

Safety measurement is historically devoted to understanding problems and mistakes. This kind of measurement is known as Safety I.³⁵ Several people we interviewed from different stakeholder communities argued that the rate at which an AV does something correctly—the successes—should also be counted. As one person put it:

³⁵ Erik Hollnagel, Robert L. Wears, and Jeffrey Braithwaite, *From Safety-I to Safety-II*, University of Southern Denmark, University of Florida, and Macquarie University, white paper, 2015.

[Safety] needs to be demonstrated. One of the beauties of the AV is its ability to see in infrared, 300 meters ahead, in 360 degrees, with lidar and excellent camera technology. That needs to be part of the demonstration. You could avoid getting rear-ended if you have a rear-facing camera and the car can get you out of the way. A human couldn't do that. So AVs need to be better, but we need to demonstrate these advances of the technology to the public. We need to demonstrate what these AVs could do that humans could never do. Many people don't think about that point.

This concept of measuring success is known as Safety II.³⁶ Originating in the complex environment of the health care system and still nascent, the Safety II concept has three main strengths: Successes occur frequently, Safety II reflects the AV's ability to adapt and respond to conditions, and Safety II shifts goals from the specific "don't crash" to the broader "drive safely" and from "avoid that something goes wrong" to "ensure that everything goes right."³⁷ As AVs and AV safety mature, successful execution of driving tasks might merit emphasis on success together with mistakes made by AVs when considering evidence of safety. This prospect harkens back to the need for roadmanship measures; measures described in Table 3.2 theoretically could be used to measure all the things that went right rather than only the things that went wrong.

Validity and Feasibility of Measures for Assessing Safety

The history of measures serving as approaches to assessing motor vehicle safety goes back almost as far as the history of motor vehicles themselves. Although refinement of measures is necessary, the public and regulators already know and understand them.

Measurements generally have high validity in their ability to demonstrate evidence of safety. Assuming sufficient levels of exposure, event-based and exposure-based measures are robust in that they can be seen as representing how the AV will probably comport itself in the conditions that it is most likely to encounter. Roadmanship and other leading and lagging measures also have the major strength of conveying information about performance in technology-agnostic ways.

Despite this, the safety as a measurement approach has problems with feasibility. Chiefly, measurements require that the ADS be static during the period under measurement. But AVs are continually updating:³⁸ How indicative is yesterday's performance of tomorrow's—or even today's? An ADS's ability to be updated for multiple vehicles—with a significant percentage of the fleet changing its driving profile all at once—is healthy and one of the huge advantages of AVs over conventional vehicles. However, if the ADS is changing every day, the past week's data will not reflect performance on a given day. An analogy for human drivers might be taking

³⁶ Hollnagel, Wears, and Braithwaite, 2015.

³⁷ Hollnagel, Wears, and Braithwaite, 2015.

³⁸ Updating involves testing internal to the developer, with scenario-based simulation followed by closed-course testing before the updates are used in vehicles on public roads.

an average measure of a person's entire driving performance between ages 16 and 40 and then assuming that it represents that person's performance at age 40 rather than finding a measure that understands how the driver improved over time. And of course, not every update will improve performance, given the imperfect reality of software development. If other road users have specific expectations of AV behavior and traffic flow based on previous behavior that is then changed by an update, human drivers might be caught off guard when the AVs react differently after the update, which could result in degraded safety even if the update itself improves AV performance.

One option could be a measure for which events occur so frequently that a day's worth of data would be informative, but such a measure would likely reflect safety only weakly. Previous RAND reports described another option for dealing with this issue: using a moving average measure process that balances forgiveness and forgetfulness of past performance by weighting the recent past over the distance past.³⁹ However, a rolling measure should be restarted under two circumstances:

1. when the ODD changes
2. when measurement methods, precision, or technology change.

Either situation implies that data from a previous performance period will not reflect current performance.

Another issue with measurements for characterizing safety is that measures are deeply rooted in an ODD, as discussed throughout this report. Although ODD might not influence a measure's validity, it could influence how the measure is analyzed and understood. For example, if data about human drivers are not available for the same ODD, comparing AV measures with aggregated human measures becomes misleading.⁴⁰ Additional issues with using statistical measures, notably in their ability (or lack thereof) to provide thresholds for safety performance, are discussed in Chapter 5.

Uniformity Versus Customization of Safety Measures

How important is uniformity across the industry in matters of safety measurement? Our 2018 report spoke to the potential for a common measurement framework, which was affirmed during the research that culminated in this report. In the words of one interviewee, "There is enough experience from the last five years that there should be some common measures." Another spoke of a need for "protocols for generating measures." Also affirmed throughout the interviews was

³⁹ Fraade-Blanar et al., 2018.

⁴⁰ One option under exploration is for the AV to gather data on the surrounding human-driven vehicles with which the AV shares the roadway when the AV is in automated or manual mode. This data will create ODD-specific benchmarks of human-driving performance. Such benchmarking assumes that human-driving behavior is not affected in any way by the presence of the AV (which would taint the data) and is broadly representative of human-driving behavior in the ODD.

the continuing tension between developer reliance on undisclosed, internal, company-specific measures supporting specific development efforts and the recognition of external interests in seeing the output of meaningful measures and in comparability across developers.

In the continuing absence of commonly used measures, stakeholders outside the industry are pragmatic: They use whatever measurements available. As one interviewee put it:

[There is] no consensus on metrics. Your previous work served as our bible for the last year and a half. At this point—it would be great if there were consensus, but, in its absence, we would entertain different methods.

However, this ad hoc approach muddies any effort to understand safety across the industry overall and the safety of individual entities. The situation is complicated by a diverse industry using different use cases (and different ODDs). For example, as one interviewee stated:

There definitely is value in having a standardized approach, but we shouldn't have one that precludes the other. Some standardized methodology, possibly standardized metrics associated with ADS-operated safety. Because of variation in use cases, vehicle systems, and subsystems, there should also be latitude for individual companies to work with their stakeholders broadly—municipalities for AV shuttles, robo-taxis; Federal Highway Administration for highway applications, Federal Transit Administration for transit—and develop application- and ODD-specific measures relevant to safety.

As another said:

A standard set of measures is needed, but each manufacturer is moving forward with its own technologies, some common and some not. Measuring may mean different things to different manufacturers. There is a goal of standard measures; the challenge is in the details.

Interviewees acknowledged that this diversity of goals and circumstances requires some latitude for individual companies to work with different kinds of stakeholders to be able to develop specific measures relevant to safety.

Reporting on measures remains the most contentious area. One official expressed frustration that developers are not providing data that are relevant at state and local levels. Two other interviewees lamented:

We can't assess safety just once, since each manufacturer is doing things very differently, in both hardware and software. We've got to develop physical shared tests that are visible to the public (not just virtual).

Every company is doing internal safety benchmarking, defining safety-critical incidents. It would help us to understand how they're doing, how they measure, how they decide . . . when [they are] ready for next step.

Government officials whom we interviewed acknowledged that the public can be critical of government and industry. The possibility of a neutral third party in a type-approval model of oversight (discussed further in Chapters 4 and 7) was recognized by all kinds of stakeholders as

an option that might appeal to some but that also might not work. It is also possible that this model could generate new issues, given the persistent constraints on the amount and kinds of information that developers will share and the level of expertise needed to understand an individual developer's data.

4. Safety as a Process

In the face of uncertainties about what can be measured and what information might show evidence of AV safety, developers and other stakeholders point to processes or specific sets of steps that might be used by developers to assess safety. Processes can point to what some engineers call a *positive trust balance*. As explained in one interview:

The positive trust argument says that we'll never have enough data in the real world to "prove" [that an AV] is safer than human-driven vehicles. So, part one of the positive trust argument is "be as safe as you can be." If you test for 10 million miles with no mishaps and no changes to the system, you can say "I'm perfectly safe"—with an asterisk, with limited data. You think you're safe, but you're still just guessing. Part two, you need very aggressive field feedback—the car almost hit a person, let's figure out why. Part three, great engineering. Safe enough isn't a number, it is an argument. "As low as reasonably possible" sounds good, but you can only calculate in known unknowns.

Processes organize the activities encompassed in this characterization of how to achieve positive trust balance (including activities associated with verification and validation of the engineering) along with other activities and provide a basis for communicating them.

Processes also provide a signal about how AVs might have been engineered to handle new situations. Handling such situations is an enduring and even defining challenge that must be met to support statements about safety.⁴¹ In contrast with measurements, which apply to phenomena that have occurred, processes are prospective indicators. As one interviewee put it:

Part of the equation is being able to demonstrate you are meeting the available standards . . . , that you have a robust safety case. You need these before you can decide if you are "safe enough." The question "how safe is safe" is where things get interesting. If you are minimalist, you must meet the appropriate regulations for your product. Past that, it depends on the product, what ODD you have, what features that you're trying to get to market, and what the risk standards are compared to [what is in relevant] ISO [International Organization for Standardization] standards. There is no specific benchmark. It's the process and the argumentation that are used that is more important to me than any one number.

Processes can take different forms. Most commonly discussed for AVs—in our interviews and more broadly—are industry standards developed under the aegis of a technical-standards-setting organization, adherence to government regulation, and industry best practices. These can provide bases for the first two categories but also give scaffolding for safety culture. The process category is internally synergistic, in that processes promulgated by technical standards-setting

⁴¹ Testing (e.g., with simulation or scenarios) is limited by the ability to anticipate, which will always be incomplete.

organizations can also be encouraged or even required by government and by corporate policy. Options to demonstrate AV safety through process have been growing. As discussed in a later section, demonstrating compliance with a process is a way for developers to document the steps they are taking to promote safety.

Safety Cases as Crosscutting Presentations of Evidence for Acceptably Safe

An element of some safety-related processes, especially for software-based systems, is a formally articulated safety case.⁴² A *safety case* is a way of packaging formal and other assertions about safety (see Chapter 6) to reduce safety and other (e.g., commercial) risk that has been used in nuclear, defense, health, aviation, and other contexts for decades.⁴³ As one interviewee explained,

With the safety case–based approach, we make the assertion that we think our vehicle is proficient in a variety of matters. We think this is a useful framework and approach. We’ve seen this type of safety approach used in other industries (aviation, medications). What is the due diligence we need to perform as a company, not just with the vehicle, but as a whole organization to manage and mitigate risk?

Safety cases can serve as road maps, one reason for their use in technical standards (discussed in the next section). Safety cases emerged as “a departure from highly prescriptive safety standards” associated with “compliance with predefined objectives.”⁴⁴ Safety cases connect the behavior of a product to the process of its development.⁴⁵ An illustration is the “Safety Case Framework” published by Uber Advanced Technologies Group in July 2020 and discussed in its updated safety report published in August 2020.⁴⁶

⁴² Robin E. Bloomfield and Peter Bishop, “Safety and Assurance Cases: Past, Present and Possible Future—an Adelard Perspective,” in Chris Dale and Tom Anderson, eds., *Making Systems Safer: Proceedings of the Eighteenth Safety-Critical Systems Symposium*, Bristol, UK, 9–11th February 2010, London: Springer, 2010.

⁴³ Bloomfield and Bishop, 2010.

⁴⁴ Ewen Denney, Ganesh Pai, and Josef Pohl, “Heterogeneous Aviation Safety Cases: Integrating the Formal and the Non-Formal,” in *IEEE 17th International Conference on Engineering of Complex Computer Systems*, Paris: Institute of Electrical and Electronics Engineers (IEEE), 2012, p. 199.

⁴⁵ Bloomfield and Bishop, 2010.

⁴⁶ Uber Advanced Technologies Group, “Safety Case Framework,” website, undated; Uber Advanced Technologies Group, *A Principled Approach to Safety: Safety Report 2020*, August 2020; Uber ATG Safety Team, “Uber ATG Releases 2020 Safety Report,” *Medium*, August 28, 2020.

Safety as Indicated by Compliance with Technical Standards or Best Practices

Compliance with standards is a process often used for contexts in which formal regulation is limited or production is global.⁴⁷ Standards are developed by representatives of companies and others (e.g., researchers) who find common cause in establishing parameters for performance and design and for mechanisms that allow interoperability or consistency in training. One interviewee explained the need as follows:

[As a developer,] I should show what math I used and why I thought it was reasonable but, in those equations, plug in assumptions about the behavior of the rest of the actors—vehicles, pedestrians, etc. It would be great if I made assumptions based on standards everyone agreed on. If those assumptions prove wrong, they could revise the standard. [We need a] standard idea of what is okay to assume about how other actors act.

Standards for conventional motor vehicles have set a high bar—what an industry analyst interviewee referred to as “auto-grade.”

Standards help an industry to function and build in best practices.⁴⁸ As an interviewee put it,

[With] standards, things need to be mature enough that most of the short-term confidentiality things are already done: The products have been discussed in public, in academic papers. In general, people are not going to participate in standards organizations unless they’re willing to share information, time, and data. That information is only valuable if there are trade secrets—but when two companies get together, there almost never is as much of an advantage as companies think there is. You tend to not get much of a diversity of solutions. . . . Everyone is working with similar basic hardware and the same physics. Everyone converges to the best solutions over time.

In part because standards are shaped by competitors, they can be set at a high level to allow companies to pursue individual, idiosyncratic designs and approaches. This is especially likely for performance standards, which are the predominant kind relating to AVs and which accommodate different choices of technology and designs to meet performance goals. Standards have always taken years to develop.⁴⁹ The challenge of achieving agreement in an international

⁴⁷ Khalid Nadvi, “Global Standards, Global Governance, and the Organization of Global Value Chains,” *Journal of Economic Geography*, Vol. 8, 2008; Petra Christmann and Glen Taylor, *Firm Self-Regulation Through International Certifiable Standards: Determinants of Symbolic Versus Substantive Implementation*, paper submitted to the first annual Conference on Institutional Mechanisms for Industry Self-Regulation, Dartmouth University, Hanover, N.H., November 14, 2005. For AVs, developers worldwide look to standards developed through ISO, the United Nations Economic Commission for Europe (UNECE), the British Standards Institution (BSI), and other country-based standards-setting organizations.

⁴⁸ Nadvi, 2008.

⁴⁹ The long time involved is also an argument used against premature regulation in a context of technologies evolving comparatively quickly.

industry—with developers demonstrating national differences in preferences for how prescriptive a standard should be—also adds delay.

Since *Measuring Automated Vehicle Safety: Forging a Framework* was published,⁵⁰ the number and variety of AV standards have grown (Appendix C lists notable examples), as have standards concerned more generally with safety of systems using artificial intelligence. In our prior research, one international standard dominated discussion: ISO 26262, first published in 2011. This standard is concerned with electrical and electronic faults and failures as a source of safety problems and calls for a safety case.⁵¹ It is broadly acknowledged as necessary but not sufficient for AV safety. A complementary standard, ISO 21448, anticipated in our previous report and first published in January 2019, focuses on intended motor vehicle functionality when complex sensors and algorithms are responsible for situational awareness.⁵² In 2019, a group of AV developers shone a spotlight on the growing corpus of technical standards. The group’s white paper provides a catalogue of standards and best practices and combines cybersecurity with safety.⁵³ (Our previous report mentions the unfortunate separation of most conversations about each of those concerns.⁵⁴) Here, we discuss notable examples of recent relevant standards-development activity.

A very different kind of standard, UL4600, was published in April 2020. It is an outlier in the AV arena in that it was developed through Underwriters Laboratory,⁵⁵ which has broad recognition for its support of safety (including in connection with consumer protection), and its development engaged a broad array of stakeholders rather than a narrower, technically oriented group.⁵⁶ Safety cases are central to UL4600, which attempts to encompass what developers do in support of narrower technical standards and allows for custom selection of activities associated with risk assessment, automation and associated engineering, and life-cycle considerations.

⁵⁰ Fraade-Blanar et al., 2018.

⁵¹ Automotive IQ, “Car Safety: History and Requirements of ISO 26262,” webpage, June 29, 2016; ISO, “Road Vehicles—Functional Safety—Part 1: Vocabulary,” ISO 26262-1:2011, 2011.

⁵² Junko Yoshida, “AV Safety Ventures Beyond ISO 26262,” *EE Times*, March 5, 2019; ISO, “Road Vehicles—Safety of the Intended Functionality,” ISO/PAS (Publicly Available Specification) 21448:2019, 2019.

⁵³ Aptiv Services, Audi, Bayrische Motoren Werke (BMW), Beijing Baidu Netcom Science Technology, Continental Teves, Daimler, Fiat Chrysler Automobiles (FCA), HERE Global B.V., Infineon Technologies, Intel, and Volkswagen, *Safety First for Automated Driving*, 2019.

⁵⁴ Fraade-Blanar et al., 2018. ISO is also working toward a standard, starting with a technical report, that combines safety and cybersecurity for motor vehicles and was informed by Aptiv et al., 2019. See ISO, “Road Vehicles—Safety and Cybersecurity for Automated Driving Systems—Design, Verification and Validation Methods,” ISO/CD TR 4804, undated.

⁵⁵ Underwriters Laboratories, “About Us,” webpage, undated-a.

⁵⁶ Junko Yoshida, “Safe Autonomy: UL 4600 and How It Grew,” *EE Times*, April 2, 2020; Underwriters Laboratories Standards, “Standard for Evaluation of Autonomous Products,” Standard 4600, Edition 1, April 1, 2020.

The IEEE P2846 working group (with leaders from Intel, Waymo, and Uber Advanced Technologies Group)⁵⁷ launched in January 2020 with the aim of forging a standard for “A Formal Model for Safety Considerations in Automated Vehicle Decision Making.”⁵⁸ This standard will facilitate using additional specialized software as a check on decisions made by the primary ADS software, and it will also support a potential for multiple developers to use common software modules.

In 2019, the National Institute of Standards and Technology began to build on its long-standing work on cyber-physical systems by convening industry and other experts to explore additional areas where standards might be developed for AVs.⁵⁹ After a general workshop,⁶⁰ another convening effort was launched in 2020 focusing on ODDs and uses of scenario-based testing.

Building a Shared Terminology Through Standards-Setting

Processes build from or shape common language. New technologies or products begin with labels and other terms chosen by their developers, which can become another arena for competition.⁶¹ Convergence to common language demonstrates maturation and facilitates comparability and comprehension by observers. For example, the 2016–2018 designation of

⁵⁷ IEEE, “WG: VT/VTS/AV Decision Making,” webpage, undated. IEEE also has other AV-related project working groups: P2040 on general requirements for fully automated driving on public roads, P2040.1 on taxonomy and definitions for connected and automated vehicles, P2040.2 on recommended practice for multi-input-based decisionmaking of automated vehicles driving on public roads, and P2040.3 on recommended practice for permitting automated vehicles to drive on public roads (Institute of Electrical and Electronics Engineers Standards Association, “P2040—Standard for General Requirements for Fully Automated Vehicles Driving on Public Roads,” webpage, undated-a).

⁵⁸ Institute of Electrical and Electronics Engineers Standards Association, “P2846: A Formal Model for Safety Considerations in Automated Vehicle Decision Making,” webpage, undated-b; IEEE, undated. The IEEE working group explains its motivation as follows:

Industry implementers creating “Safe By Design” automated vehicles, as well as government and independent assessors, need a metric to assess whether an automated vehicle is driving safely according to the agreed-upon balance between safety and practicability that is at the heart of driving in the real world. Without a formal model for automated vehicle decisionmaking, industry will not know how safe is safe enough, and government will not have a tool to define what safe driving means. . . . This standard defines a technology-neutral formal model, parameterized so that the balance between safety and utility of automated vehicle decisionmaking may be adjusted to reflect different cultural and other differences in what it means to “drive safely.” The value of a technology-neutral model is that it is compatible with not only any kind of planning function (rules based, or machine learning) but is flexible enough to be integrated into any automated driving system (ADS) architecture. (American National Standards Institute, “Project Initiation Notification System [PINS],” *ANSI Standards Action*, Vol. 51, No. 27, July 30, 2020, p. 48)

⁵⁹ National Institute of Standards and Technology, “Cyber-Physical Systems,” webpage, undated.

⁶⁰ Edward R. Griffor, Christopher Greer, and David A. Wollman, *Workshop Report: Consensus Safety Measurement for Automated Driving System-Equipped Vehicles*, Gaithersburg, Md.: National Institute of Standards and Technology, September 23, 2019.

⁶¹ Tesla’s use of Autopilot presents an extreme case. Tesla, “Future of Driving,” webpage, undated.

capability levels for AVs by SAE was adopted around the world and across stakeholders.⁶² The 2020 alliance of consumer and industry groups supporting common nomenclature for Advanced Driver Assistance Systems (ADAS) represents a push for common public-facing terms.⁶³ Similarly, the 2020 developer-based Automated Vehicle Safety Consortium (AVSC) publication of a common approach to describing ODDs supported the comparison of different AVs.⁶⁴ Technical standards-setting activities typically begin with agreement on terminology.

Safety as Indicated by Compliance with Federal, State, or Local Regulations

The standards proliferation we have outlined is consistent with a dynamic situation in which technology is evolving quickly, mechanisms for self-regulation—commonly associated with processes—loom large, and government regulation is limited.⁶⁵ Although the interviews conducted for this project captured contrasting stakeholder views of the limited scope of extant regulatory portfolios, several people who are or were in government acknowledged the limits of understanding of AVs in government at any level. The following quotation is illustrative:

In the new space that we are in right now with AVs, I think you have to go with voluntary standards. If you lock it up in a government regulation, it takes forever to get it in and may be based on old science. Until we figure out the categories, voluntary standards are more nimble and responsive.

Regulation guides AV development at each level of government. At the local level, municipalities and counties own roads and can regulate the kinds of vehicles that can operate on them; these local governments manage traffic and enforcement of traffic laws. The local and state levels are where one can find commitments to Vision Zero, focusing on the traffic system and not just individual vehicles (as noted in Chapter 1). At the state level, governments control and manage roads and are responsible for testing and licensing of drivers, regulation of insurance companies (and establishment of requirements for drivers to have insurance), traffic laws, and rules of the road. State and local governments are gatekeepers for AV testing on public roads in their jurisdictions. States have been developing frameworks involving laws, executive orders, regulations, special studies, and baseline standards and information collection addressing not

⁶² SAE International, 2018.

⁶³ American Automobile Association, Consumer Reports, J. D. Power, National Safety Council, and SAE International, “Clearing the Confusion: Recommended Common Naming for Advanced Driver Assistance Technologies,” webpage, undated.

⁶⁴ AVSC, *AVSC Best Practice for Describing an Operational Design Domain: Conceptual Framework and Lexicon*, Warrendale, Pa.: SAE Industry Technologies Consortia, AVSC00002202004, 2020. Fraade-Blanar et al. (2018) recommended the development and use of a common approach to describing ODDs.

⁶⁵ Christmann and Taylor, 2005.

only testing but also associated issues beginning with safety and extending to broader issues, such as infrastructure and land-use planning.⁶⁶

At the federal level, Federal Motor Vehicle Safety Standards (FMVSS) are intended to “prevent or reduce vehicle crashes.”⁶⁷ The expectation of compliance with FMVSS guides vehicle design and engineering, and the federal government can recall vehicles already on the road in the event of demonstrated safety shortcomings. People familiar with AV development (including all of the stakeholders interviewed for this project) understand that today’s FMVSS have limited applicability to ADS and that they contribute to the safety associated with the physical features of the vehicle (crashworthiness and occupant protection) rather than how the vehicle operates.⁶⁸ Most stakeholders consulted for this project were skeptical that federal safety standards for automated driving systems could be established any time soon. One government official described the situation as follows:

It is difficult to define a uniform standard for the industry because it is still premature. There are different systems, different use cases, different environments, different ODDs where vehicles are being tested. There is some distance to go before there is enough data in all of those different categories that are then cross-referenced to generate the kind of uniformity needed for “safe enough.” A premature imposition of operational standards could be counterproductive or ineffective. . . . It is better to not establish performance measures up front—it is better if that happens through a sort of evolutionary process.

Even before the advent of the AVs now being tested, the development of FMVSS has been slow; the timetable built into the U.S. rulemaking process was noted by a variety of stakeholders.⁶⁹ The federal government also attends to concerns related to interstate transportation and associated interoperability concerns.

Stakeholders we interviewed generally acknowledged that developing FMVSS or other approaches to safety regulation is complicated by the fact that AV technologies have not stabilized. The dependence on software adds to the challenge for regulation relating to AV safety, as has been seen in other industries (e.g., aviation, medical devices). That challenge is

⁶⁶ National Governors Association Center for Best Practices, *State Public Safety and Autonomous Vehicle Technology*, Washington, D.C.: National Governors Association, 2018. Vermont, for example, is one of the states to most recently spell out the scope of authority for that state and its municipalities (Vermont Agency of Transportation, Policy, Planning, and Intermodal Development Division, *Vermont Automated Vehicle Testing Permit: Guidance and Application*, Barre, Vt.: Vermont Agency of Transportation, April 24, 2020).

⁶⁷ NHTSA, “Regulations,” webpage, undated-c.

⁶⁸ Some of the requirements for physical features (e.g., steering wheels) have been questioned by developers designing for a future without human drivers, and corresponding flexibility has begun to emerge. The first exemption to specific FMVSS was granted by NHTSA to Nuro for a low-speed vehicle that will never carry people. (NHTSA, “Nuro, Inc.: Grant of Temporary Exemption for a Low-Speed Vehicle with an Automated Driving System,” Docket No. NHTSA-2019-0017, *Federal Register*, Vol. 85, No. 7826, February 11, 2020a)

⁶⁹ People we interviewed also observed that a certain amount of penetration of the fleet (perhaps 20 to 30 percent) is needed before there are enough data to do the analysis required for regulation.

compounded by the expectation of an ongoing series of updates. The following quotation is representative of comments collected from the industry and also others.

A lot of an AV is predicated on software, but there is no way to regulate software. It's more about processes. The expectation that human developers can develop error-free code is unrealistic. The expectation should not be for zero bugs but to find and fix bugs quickly.

The rise of partial automation through ADAS has already demonstrated the challenge of blending software engineering and mechanical engineering, a challenge that is amplified in AV development.

Government regulations and other engagements shape the playing field by defining minimum performance and reporting requirements.⁷⁰ Government officials we interviewed joined industry and other stakeholders in describing government's role as providing a floor for safety. Industry executives asserted that best practice will be at least comparable to that floor. For the public, a minimum expectation can provide assurance. This floor also limits the extent to which government officials need to contend with the diversity of developer approaches, which can be a challenge during this young phase of the industry when there are more competitors than are expected to be in operation after the industry matures.

At all levels, government's degree of activity can vary in terms of consulting with developers and being vigilant in enforcing the mechanisms that government has to work with. Government officials we interviewed acknowledged their limited understanding of the technology. Nonetheless, they are open to working with developers to learn about and inform plans (e.g., at the local level to discuss how people behave on or near roadways and other aspects of the ODD).⁷¹ The research team heard anecdotes about both good and deficient communication.

Safety as Indicated by Corporate Safety Culture

What do AV developers say, what do they do, and how do they do it? These are signals of *risk preferences*—how risk-averse or risk-seeking an organization is and how it weights public interest compared with self-interest in effecting a balance. These characteristics are also

⁷⁰ The U.S. system features self-certification for compliance with FMVSS; unlike in other countries, there is no government approval prior to commercialization of motor vehicles. The National Transportation Safety Board recommended in November 2019 “that NHTSA require entities wishing to test a developmental automated driving system on public roads to submit safety self-assessment plans before being allowed to begin or continue testing and that NHTSA should review the plans to ensure they include appropriate safeguards” (National Transportation Safety Board, “‘Inadequate Safety Culture’ Contributed to Uber Automated Test Vehicle Crash—NTSB Calls for Federal Review Process for Automated Vehicle Testing on Public Roads,” news release, November 19, 2019b).

⁷¹ *Vermont Automated Vehicle Testing Permit* guidance advises that, “[w]hile Applicants are not responsible for seeking pre-approval from a municipality, they may find it beneficial to work directly with a municipality in advance of submitting a permit application when town highways are a critical part of a testing plan and are certainly free to do so” (Vermont Agency of Transportation, 2020, p. 7).

indicative of what is referred to as *safety culture*. Two quotations from the project interviews are representative:

If people see [that] companies are doing their best, they will be much more forgiving than if corners are being cut to put profit first. People will see through it.

We can't evaluate the technologies, so we can only evaluate and trust the companies.

Stakeholders we interviewed frequently pointed to a developer's safety culture as an indicator of how it approaches safety-critical processes. Safety culture might be particularly helpful when specific processes break down or become disconnected from measures (see Chapter 3).

AV development features unique circumstances that can be linked to safety culture. Notably, these include treatment of *safety drivers*, the humans overseeing the ADS as it is tested on the road and making decisions about disengagement. Aspects to consider include safety-driver status (e.g., contractor or employee), empowerment, and training. Furthermore, AVs generate massive amounts of data, and safety culture guides choices about using those data for continual improvement.

Because the details of most of what a developer does—from its work with safety drivers to detailed engineering decisions or actions—are proprietary, there seems to be agreement across stakeholders on the value of higher-level, enterprisewide culture. Safety culture informs both substantive contributions to AV safety, per se, and signaling about the overall culture of the enterprise, contributing to public engagement. Because AVs are novel and have attracted huge investments, developers need ways to communicate their sincere concern for safety as they pursue profit. Evidence of a safety culture can help. Three representative quotations from the project interviews are as follows:

“Safe enough” is not a number—it has to be more of a safety-culture thing, all this stuff tied together. If you hit a number, [it should mean that] you didn't get lucky but used an intentional process; if you guessed wrong, you will fix it.

Until there are definitions of metrics, the only thing we can do is reveal the people behind the companies doing the work. This is our communication of safety.

What are the early warning signs that something will be a problem? Making companies share their safety philosophies sets up the possibility to give some early indication that they should be scrutinized or do not deserve trust. If a company is making impossible claims, or intentionally misleading claims, I would shut them down. So much of this revolves around the necessity that AV companies act on good faith. If there are failures at this step, that is an indicator that the company may be a bad actor.

Safety culture interacts with best practices associated with software development and other aspects of AV development. The technical standards we have outlined address some aspects of AV safety; more-general guidance exists for developing software that is supportive of safety, cybersecurity, and/or dependability.⁷² Best practices are under development by the relatively new AVSC. Unlike standards-developing organizations, which tend to convene large committees that work toward consensus, AVSC is small enough to come to agreement faster than often happens with technical standards, which could be bootstrapped by the best practices that AVSC develops.⁷³

Insights can be gleaned from other domains. For example, the Federal Aviation Administration established the Safety Management System, a set of processes for managing system safety and safety management through a “structured process that obligates organizations to manage safety with the same level of priority that other core business processes are managed.”⁷⁴ This system has four principal components (safety policy, safety risk management, safety assurance, and safety promotion),⁷⁵ and it calls for an independent safety organization within the enterprise that is not part of development but partners with the development. The Safety Management System combines principles, practices, and oversight that could be valuable in the AV industry,⁷⁶ and our team’s consultations indicate that it has begun to be used in that context.

People we interviewed remarked on the value of having an independent safety department, personnel with safety-related training, use of data to drive decisionmaking, and engagement from the chief executive officer (and reporting of safety management to the chief executive officer)—stakeholder interviews showed broad agreement that safety starts at the top. Some referred to the importance of an ethos that could be characterized by such statements as “breathing safety,” “every person from the director to the trash collector feeling the safety,” or “having it [safety] in their bones.” Bryant Walker Smith has summed up the situation by stating, “Safety is a marriage, not just a wedding; a lifelong commitment rather than a one-time event.”⁷⁷

⁷² “A safety culture and the processes that support it need to be accompanied by the best technical practices in order to achieve dependability” (National Research Council, *Software for Dependable Systems: Sufficient Evidence?* Washington, D.C.: National Academies Press, 2007, p. 49).

⁷³ The involvement of SAE, which is a traditional standards-setting organization, in AVSC would facilitate such a transition.

⁷⁴ Federal Aviation Administration, “Safety Management System (SMS),” webpage, June 21, 2019.

⁷⁵ Federal Aviation Administration, “Safety Management System (SMS): Components,” webpage, September 11, 2017.

⁷⁶ Accordingly, the National Transportation Safety Board recommended it to the Uber Advanced Technologies Group, which has embraced it (National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian: Tempe, Arizona, March 18, 2018*, Washington, D.C., November 19, 2019a; Uber Advanced Technologies Group, 2020).

⁷⁷ Bryant Walker Smith, “How Reporters Can Evaluate Automated Driving Announcements,” *Journal of Law and Mobility*, April 19, 2020.

Safety culture can be expressed at the industry and the company levels. Interviewees with broad knowledge of safety, including how it is handled in a variety of industries, noted that, in transportation industries with comparatively few players (such as aviation and rail), safety culture extended to a willingness to share information about problems.⁷⁸ In the AV domain, however, companies appear more willing to share philosophies and principles than to share data.

⁷⁸ Such industries also evidence different regulatory histories and risks of catastrophe, neither of which diminish the benefit that has been gained within the industries from sharing information.

5. Safety as a Threshold

A threshold can establish the level of safety that an AV must achieve to be considered acceptably safe. Many technologies lack strictly defined levels of acceptable risk. What differentiates AVs from, as one interviewee put forth, a toaster, is the possibility of death to users and bystanders. For AVs, a threshold acts as a minimum requirement or a safety floor that must be met and can be exceeded. This floor could reflect what is legal and feasible, or it could demonstrate that AVs are not increasing risk or harm on the roadway.

Thresholds come in multiple forms. They draw from rather than act as alternatives to measures and processes. Thresholds can be quantitative, such as reaching a fatal crash rate that is a given percentage lower than the human-driver crash rate or achieving a prescribed fatality rate.⁷⁹ Thresholds can be binary, such as passing a driving test.⁸⁰ Alternatively, thresholds can combine quantitative and qualitative evidence into judgments about whether the technology is as safe as possible.⁸¹

One challenge to identifying AV safety thresholds is that thresholds are traditionally set after a given technology has been in use for a time, whereas AV technology remains a work in progress. As one interviewee observed, “We have been stuck in reactive mode for 100 years. Now we switch to being proactive.” Being proactive is tricky because (1) the newness of the technology means that existing thresholds (e.g., those embodied in FMVSS and driver’s license tests) and methods of assessing thresholds (e.g., a driving instructor inside the vehicle, assessing the human driver) no longer guarantee a floor and (2) thresholds should evolve as the technology evolves. Thus, meeting a threshold is not a one-time achievement.

“The definition [of acceptably safe],” explained one interviewee, “will be iterative and dynamic depending on how the technology advances.” Having evolving, multiple thresholds sidesteps concerns that the field is not ready for a threshold. This could also avoid a peril, acknowledged by several people we interviewed, that (in the words of one) “premature imposition of operational and performance standards (and thereby thresholds) could either be counterproductive or ineffective.” Evolving, multiple thresholds can complement and reinforce each other. These thresholds exist along several dimensions.

⁷⁹ These issues are discussed in later sections of this chapter. See “Safety as Achieving a Threshold Based on Human Driving Performance” and “Safety as Achieving a Threshold Based on an Absolute Safety Goal.”

⁸⁰ See the section in this chapter titled “Safety as Achieving a Threshold Based on ADS Technology Potential.”

⁸¹ This material is covered in greater detail later in this chapter. See the section titled “Safety as Achieving a Threshold Based on ADS Technology Potential” and Box 5.1 at the end of this chapter.

First, as AV technology advances, the threshold that the technology must meet also advances.⁸² Although the form of the threshold might not change, the level of the threshold will. For example, developers might argue that AVs should achieve a crash rate in year one that is 50 percent lower than the human-driver crash rate. But in year 10, the crash rate should be 90 percent lower. If AVs are safe enough to “deploy when they can save more lives than humans can,” explained one interviewee, “it’s not reasonable to stop work given the duty of care to continue to improve technology until it reaches some critical level of diminishing returns.”

Second, new thresholds are needed every time the AV’s ODD expands. Herein, the form and level of the threshold could change. For example, the evidence used to show capabilities when driving five miles from home base might be different from the evidence used to show capabilities when that range is expanded to 50 miles from home base.

Third, as AV developers make operational decisions, new thresholds must be met before enacting said decisions, both before and after commercial deployment. The form of the threshold might change because different types of evidence are needed to speak to different operational decisions. For example, different thresholds would be required for decisions to remove safety drivers, allow access to the AV via a publicly available app, or allow nonlicensed drivers (e.g., children or the disabled) to ride alone in the AV. In the words of one interviewee:

“Safe enough” is a continuum. You won’t go straight from “no AV” to “commercial deployment” (which I think of as a large fleet driving millions of miles a year). This won’t happen in one leap. So, what does “safe enough” mean—safe enough for what? When will cars pull the “safety driver” out of the car? . . . Then you will elevate the difficulty as you go and advance. It’s not zero to commercial deployment.

Fourth, ADS hardware and software updates will have to show that they meet all existing thresholds. One technologist interviewee wondered “what happens when [an] ADS ‘wears out’ over time” because safety performance will change as “mechanical parts are subject to degraded function, and logical parts of software, including algorithms, evolve.” AVs must meet thresholds not only when they first enter commercial service but also in an ongoing manner. The former has been traditional for human-driven vehicles,⁸³ but the latter is important for AVs given the possibility for ongoing over-the-air updates.⁸⁴ With AVs, explained one interviewee, there is a “shifting from ‘design once, test hard, deployment, and never touch.’ That is a dead paradigm. We must monitor whether the assumptions are right post-deployment.” That characterization reflects the nature of AD as a software-based product. It is typical for software products to be

⁸² In this dimension, the threshold was described by interviewees as “progressive,” “evolving,” “retesting,” and “a moving target.”

⁸³ Here, we refer to thresholds beyond annual inspections, which some states do require for human-driven vehicles.

⁸⁴ Some current ADAS and systems at the lower levels of the SAE scale (levels 1 and 2) already use over-the-air updates at this time.

released after they achieve thresholds internal to the developer and then to be updated regularly as problems or opportunities for improvement emerge. As one interviewee put it:

No one [in industry] will wait for absolute perfection across all conditions. Once it is good enough, go into pilot deployment, keep getting over-the-air updates and mods. It must be a continuous process.

Other dimensions could include evolution of customer or societal expectations of technology and improvement in safety assessment methods (e.g., technology that allows for greater precision). Multiple paths exist to enforce thresholds: federal government recalls, state and local policies and permitting, liability law, and consumer demand and behavior, to name a few. Thresholds must be technology-agnostic if they are used for purposes beyond within-company benchmarking.⁸⁵

In this chapter, we discuss three types of thresholds: those rooted in human-driver performance, those rooted in ADS technology, and those based on absolute numbers. The threshold approach contains two components: the decision of what is safe enough or acceptably safe and a determination of whether those targets have been met and how AVs will continue to do so as they are developed. Communication of thresholds is discussed in Chapter 6.

Safety as Achieving a Threshold Based on Human Driving Performance

Every stakeholder we interviewed recognized the inevitability of comparisons between AV and human safety performance and the use of the latter as a threshold for the former. Some found this natural and useful for communicating safety to the general public. However, frustration was also expressed regarding, among other things, how the comparison is done and the expectation that AVs at maturity will behave very differently from human drivers.

Thresholds Based on Human Driving Performance: The Concept

The human driver—either the average human driver or the safe human driver—was frequently cited by interviewees as a basis for a safety threshold of AVs (for example, making “comparisons to human drivers is necessary, at least in the short term”; “comparing to the baseline of a human driver is a good idea”; and an “AV needs to have the same standard as a human.” This seeming imperative to compare AVs with human drivers is echoed in the popular press.⁸⁶

⁸⁵ Fraade-Blanar et al., 2018.

⁸⁶ Lance Eliot, “Essential Stats for Justifying and Comparing Self-Driving Cars to Humans at the Wheel,” *Forbes*, May 30, 2019; Emily Stewart, “Self-Driving Cars Have to Be Safer Than Regular Cars. The Question Is How Much,” *Vox*, May 17, 2019.

Using human drivers as a basis for an AV safety approach resonates widely for three reasons: it reflects the level of safety already tolerated by the general public;⁸⁷ it is consistent with existing government regulations; and it adapts a concept with which people are familiar to a new and complex technology.⁸⁸ Furthermore, because the ADS can be (and sometimes is) anthropomorphized as a robot driver, it is natural to want an ADS to be at least as good as a human driver. As one interviewee said, “I prefer [comparisons with] human drivers. There is some consensus among companies that this is the right comparison.” Or in the words of another:

Since we are at the place of replacing human decisionmaking with machine decisionmaking, but we don’t know logically or medically how we make decisions. . . . “[G]ood enough” is the best job a human driver does.

But what kind of human driver, specifically, should be the basis for the threshold? Responses coalesced around two options: the average human driver and the safe human driver. The average human driver was most heavily discussed as an unsatisfying option along the following lines:

- The concept of an average driver is ill-defined.
- Average drivers “still have a lot of accidents on the roadway, so that isn’t a good benchmark for comparison.”
- Safer-than-average human drivers might be the first step but not the comparison metric for the longer term.
- Many expressed the goal of being not just safer than the average human driver but “dramatically better.”

One reason an average human driver might be unsatisfying as a comparison metric is that, although key statistics measure risk on the population level, people experience risk at the individual level. To illustrate, the conventional vehicle fatal-crash rate is measured per 100 million VMT (1.1 fatalities per 100 million VMT in 2019⁸⁹). But Americans drive an average of around 10,000 miles annually.⁹⁰ Over a lifetime, no one will drive 100 million vehicle miles. So, the average human-driving rate, which makes sense at the national or state fleet level, might fail to translate to individuals because it tells us relatively little about the odds of an individual driver crashing.

Additionally, most drivers believe they are safer than an average driver. Oddly, simple analysis shows that they are right. In terms of any crash, from fatalities to solely property

⁸⁷ Fewtrell and Bartram, 2001.

⁸⁸ Wishart et al., 2020. This option is somewhat predicated on the assumption that, as one interviewee put it, human-driven rides are most likely the type of transit that AVs will replace. AVs could also replace rides that would otherwise be on public transit, walking, biking, or other modes.

⁸⁹ NHTSA, “Early Estimates of 2019 Motor Vehicle Traffic Data Show Reduced Fatalities for Third Consecutive Year,” press release, May 5, 2020c.

⁹⁰ National Household Travel Survey, *Developing a Best Estimate of Annual Vehicle Mileage for 2017 NHTS Vehicles*, 2017.

damage, 6,452,000 crashes were reported to the police in 2017.⁹¹ There were 225,346,257 licensed drivers,⁹² so the mean crash rate nationally was 0.03 crashes per licensed driver. The median number of drivers—the vast majority—did not crash in 2017 and are therefore safer than the average (mean) crash rate (because they did not have 0.03 or more crashes).⁹³ Although a trick of math,⁹⁴ this analysis suggests that measurements to identify the average human driver fail to represent the average human driver’s lived experience. Because human nature can tolerate human error—those who are in accidents can make up some excuse about what happened—why should individuals accept as a benchmark a driver that is less safe than themselves?

There is also an intuitive argument that most human drivers are better than average. The ability to drive safely is not likely to be evenly distributed in the population, and a driver who engages in one risky driving behavior is likely to engage in others.⁹⁵ The majority of individuals are licensed, do not drive while intoxicated, do not constantly text while driving, etc., and even common behaviors, such as self-reported running of red lights, are not reported by the majority of drivers.⁹⁶ Crashing once has been found to be statistically associated with crashing again.⁹⁷ One study reported that in Louisiana, 5 percent of licensed drivers were at fault for 34 percent of accidents.⁹⁸ Although we know how to define excessively unsafe driving behavior, we have not historically needed to define superlatively safe behavior. Consequently, the distribution of driving ability is likely to have a strong negative skew, such that the mean is lower than the

⁹¹ NHTSA, “Police-Reported Motor Vehicle Traffic Crashes in 2017,” Traffic Safety Facts Research Note, July 2019a.

⁹² Federal Highway Administration, Policy and Governmental Affairs, Office of Highway Policy Information, “Highway Statistics Series: Highway Statistics 2017,” webpage, December 2018.

⁹³ If we assume that all the police-reported crashes involved single vehicles, then 6,452,000 drivers were involved in crashes. Even if we make the extreme assumption that all crashes were four-vehicle crashes, only 25,808,000 drivers (less than 12 percent of drivers) would have been involved in a crash.

⁹⁴ The “trick” is in comparing mean with median and a population decimal rate with an individual’s whole number.

⁹⁵ Insurance Institute for Highway Safety and Highway Loss Data Institute, “Red Light Running,” webpage, February 2020. It is likely that these behaviors translate into clusters of crash risk; an author of this report recalls previous research based on a data set of state crashes that indicated that crashes were rare but that some people crashed multiple times. For example, one individual had eight police-reported crashes in a five-year span.

⁹⁶ This can vary based on circumstances. A recent report indicated that the number of incidents of poor driving behavior has increased during the COVID-19 pandemic. (Zendrive, “Mobility Amidst Lockdown: Every Minute on the Road Is Riskier,” webpage, May 2020; Insurance Institute for Highway Safety and Highway Loss Data Institute, homepage, undated).

⁹⁷ Nick Stamatiadis and Arnold Stromberg, “Crash Involvement of Drivers with Multiple Crashes,” *Accident Analysis & Prevention*, Vol. 38, No. 3, June 2006.

⁹⁸ Subasish Das, Xiaoduan Sun, Fan Wang, and Charles Leboeuf, “Estimating Likelihood of Future Crashes for Crash-Prone Drivers,” *Journal of Traffic and Transportation Engineering*, Vol. 2, No. 3, June 2015.

median. This means that most drivers are likely to be above the mean in terms of driving skill—safe drivers are the majority.⁹⁹

A better-than-average, or safe, human driver is a preferable alternative. This is partly because the general public will not accept from AVs what it accepts from human drivers.

“As safe as a human driver” is not good enough. We need only to look at how people react to airline crashes or to defects. Consumers perceive risk differently when they aren’t in control. Even if they have high expectations of safety for themselves, it can be higher when they aren’t in control. . . . You have to absolutely get rid of the “average driver” metric. You need to be substantially better for broad consumer acceptance. Consumers will not accept the failure rate of the average human driver; you need to level up.

Another factor is that technology has the potential to be so much safer.

There are some advantages. The human-driver safety statistics are heterogeneous, and they will change in different environments. Do you compare to “average drivers” not under the influence, not on cell phones? This might be a good minimum or first bar, but once AVs become state of the art, they will be expected to perform better than that, and then the human-driver comparison will no longer be appropriate.

Or as another interviewee put it, the general goal is that “‘safe enough’ is looking at the data and seeing that vehicles can perform on par with the best of what we have on the road.”

A few definitions of safe drivers exist. We list these here with quotations from interviewees:

- Pick a percentage or numerical value better than average humans: “10, 20, 30 percent,” “a degree or order of magnitude better than humans,” “95 percent.”
- Pick a specific human profile: for example, “optimally safe would be a 37-year-old-mother, probably.” Or “a trained professional driver, who would have a lot of experience, whose license is subject to review.”
- Calculate an idealized crash rate: Use a subset of crashes, excluding crashes involving unsafe driver behavior (e.g., distraction, fatigue, drunk driving) and narrow in on the driving behavior of a sober, informed, alert, engaged driver.

Each of these three options is reasonable as a conceptual representation of a safe human driver. The first is simplest to determine. The second requires agreeing on who should serve as a model, but, with agreement, the concept should be easily communicated. The third requires an excellent, detailed set of data and broad assumptions but is appealing in that AVs will not make the same mistakes that humans do. These are practical considerations for deciding on a possible threshold, recognizing that whether AVs should be as safe as human drivers or safer and that, if the latter, how much safer is as much a matter of philosophy as of science.

⁹⁹ There is an illustrative analogy to this argument: Consider that most people have an above-average number of arms. This is because it is exceptionally rare that anyone has more than two arms, but much less rare that a person has one or zero arms. As a result, the worldwide mean number of arms per person is likely to be around 1.99 . . . , even though most people have two.

Thresholds Based on Human Driving Performance: Functional Benchmarks

Defining safe human driving is tricky; measuring it is trickier. As a result, some experts “don’t believe in comparisons with human drivers. The amount of hand waving needed to make a comparison will almost make the whole thing meaningless.” Overall, the “comparison to humans is more complicated and nuanced than people think it is. There is a lot more to it. Part of it is the bad data [available about human drivers].” Measures developed for human drivers may have less applicability for AVs, and vice versa.

The first two challenges for using human performance as a benchmark for AVs—gameability and availability—center on AV data. Some automated driving systems might do better on certain statistics than others. Additionally, the differences in implementation and reporting that plague disengagement might plague other statistics, especially leading measures that are not police-reported but would have to be developer-reported. As explained by one interviewee:

Having a uniform definition but not a uniform method would be bad—simulation versus real world, for example. Then there would be comparing of methods, instead of just looking at safety. We need to have a common definition and common methods.

Clear, transparent, and universally applied definitions and methods are required, especially for leading roadmanship measures (see Chapter 3), which are still in development. Variations in technology can make such methods difficult for leading measures. But given the availability challenges of lagging measures, it would be better to make the comparison using roadmanship measures—the opportunity to do so is a motivation to continue to enhance such measures.

The third challenge of using human performance as a benchmark for AVs centers on human data: finding data from human drivers on leading and lagging events and exposure, and relating human-driver data to AV ODD. Having human-driver data would smooth over a major issue in comparing AV performance across ODDs by providing the AV version of a par determination on a golf course: The performance of an AV operating in a complex environment can be compared with the performance of an AV operating in the same environment if the threshold is whether the AV drives as well as human drivers (or safe human drivers) in that same area. But human-driver data can be hard to come by, as these quotations attest:

It is hard to get human-driver data to compare for AVs. What is the comparison and at what level? I don’t know how we compare safety with an apples-to-apples comparison.

We don’t have measurements for an average human driver, and we have no way to make those measurements relevant for AVs.

Data on most lagging measures for human drivers are available, because crashes that result in injury or in property damage above a certain cost (generally \$750–1,000) are reported to the police, but data are lacking on crashes of less severity.

Because data on most leading roadmanship measures are not available for humans,¹⁰⁰ it is difficult to use these measures to compare AV performance with that of humans.

We get data on AVs in a certain ODD. Human data isn't collected on that level. We have lagging measures, crashes, etc., but we don't have leading measures for humans. So, we compare AV performance to human data that we actually may not have much information for. . . . We have limited data on humans but absolutely no information on near misses. There's no damage, no police report, we don't ever hear about it. When cars go over multiple lanes (from the left lane, but for a right exit)—people may make this extremely risky move, but nothing [bad] may happen. So, when I talk about managing risk, we have no baseline about how risky human drivers are. All that we can safely say is that AVs will not engage in risky behavior.

The team heard in several interviews that (in the words of one) “near misses are challenging because we don't measure them for humans.”

One place that does provide human-driver data on near misses and other measures of roadmanship is naturalistic data on human-driven vehicles.¹⁰¹ Naturalistic driving data can provide otherwise-inaccessible insights into human factors specific to driver behavior, driver characteristics, pre-crash driving events, and near misses. Naturalistic data also show the generalizability of lab-based findings to real-world settings.¹⁰² But some interviewees noted the limitations of such data. Data sets generally come from academic research studies, such as Virginia Tech Transportation Institute Strategic Highway Research Program 2,¹⁰³ insurance data, or from companies that monitor commercial drivers. It is unclear how much driving behavior in these sample populations represents that of the average human-driver concept. This is particularly true for data derived from insurance and commercial-driver-monitoring companies, where people can either self-select in (e.g., particularly safe drivers looking for insurance cost reductions) or where the sample represents a particular vocation or population. Another concern is that the quality of data and definitions of events (e.g., hard braking) from naturalistic driving might not match the quality of the data that can be gathered by the AV.

¹⁰⁰ A few leading measures of roadmanship do have information available for humans that is ample, or at least ampler than for other measures. These include such measures as hard braking and reaction time (time elapsed in between a danger presenting itself and the AV reacting appropriately and safely).

¹⁰¹ Naturalistic driving data are collected from people as they drive, usually people who have agreed to participate in such data collection. Collection of information on human-driving behavior can use a range of in-vehicle data acquisition technologies, including cameras (pointed into the driving compartment to record driver and passenger behavior and pointed outward to record driving conditions), radar, accelerometers, Global Positioning System (GPS), eye-tracking technology, and alcohol sensors. (Kenneth L. Campbell, “The SHRP 2 Naturalistic Driving Study: Addressing Driver Performance and Behavior in Traffic Safety,” *TR News*, No. 282, September–October 2012)

¹⁰² Kun-Feng Wu, Jonathan Aguero-Valverde, and Paul P. Jovanis, “Using Naturalistic Driving Data to Explore the Association Between Traffic Safety–Related Events and Crash Risk at Driver Level,” *Accident Analysis & Prevention*, Vol. 72, November 2014; Engström et al., 2019.

¹⁰³ Virginia Tech Transportation Institute, *InSight Data Access*, website, undated.

Overall, there are not enough lagging-measure AV data to use as evidence of meeting a threshold based on human-driver measures, but there are not enough leading-measure human-driver data to use as evidence of meeting this threshold.

The fourth challenge is matching human event and exposure data to the AV's ODD. Thresholds are deeply and inextricably linked to ODDs. Regional differences exist in leading and lagging measures. Comparing on the state level, Mississippi had 23.12 crash fatalities per 100,000 people in 2017; New York had a rate of 5.03. But this rate is not uniform across the state; New York City's rate was 2.40. Nor are all cities equal; in contrast to New York City, Kansas City had a rate of 22.23.¹⁰⁴ As interviewees noted, "the crash rates are completely different between Mountain View [California] and Phoenix [Arizona]."

This issue of ODD variation carries over to leading measures, as expressed in the following quotations from different interviewees:

Safe driving in Kansas is different than Boston or D.C. Someone following the letter of the law, which should be safe, may impede traffic or create hazards based on the norms for different geographies. One thing that is implied about comparing AVs to human drivers is that human drivers adjust and adapt to local conditions.

[There is a] need to think about what's safe enough for the given ODD you are operating in.

Everyone hitting the same performance metric is not the best way to roll out if [the AV is] confined to specific ODDs. They don't want [company X] to hit same performance metrics as [company Y] if [company X] focuses on San Francisco and [company Y] on narrow ODDs. And vice versa."

Safe human drivers don't like to drive in Manhattan. Manhattan drivers compromise safety for the ability to get where you need to go, aggressively pulling into lane, etc.

All miles aren't created equal.

It is very complicated to actually have an appropriate human comparison. Data don't exist to get down to the ODD level. Are we including rural roads? What's in and what's out? How do you determine what the crash rate is in your specific ODD? Different from one ODD to another. . . . The ODD data for humans doesn't exist to compare it to. Comparison to a human is much more complicated and nuanced than people realize.

¹⁰⁴ NHTSA, "Traffic Safety Facts Annual Report Tables," webpage, June 30, 2020d.

These differences could become even more pronounced if data about time of day, roadway type, level of precipitation, or any other ODD dimension were available.

For fair comparison, human data must come from the same ODD as AV data—a challenge today, given the dearth of such data on people. Adding complication, environments can move inside or outside an AV’s ODD as they change permanently (e.g., where lanes are added or streets are changed from two-way to one-way) or temporarily (with construction or large events).

Exceptions to this data drought exist. As of 2020, a few localities, including San Francisco and specific areas in Maricopa County, Arizona, were monitoring human-driving behavior to generate regional leading measures of roadmanship. If an AV’s ODD perfectly matched these localities, human comparison (average, safe, or other) could be used. Consequently, because AVs develop in “islands of autonomy”—geographically specific and narrowly defined locales—they also might be most measurable in these select islands.

Safety as Achieving a Threshold Based on ADS Technology Potential

The previous section focused on human performance thresholds based on quantitative measurement. However, threshold approaches can also relate to specific kinds of performance of ADS technology. Such thresholds have been under development and are subject to debate. Thresholds rooted in ADS performance include driving tests and subjective assessments of an AV technology’s ability to fulfill its potential.

Driving Tests

A driving test, in this context, is a formalization of the idea that a threshold could be “as safe as a human driver when completing the same maneuver. If the human doesn’t crash and the AV does, the AV is not as safe.” Because the breadth of scenarios that an AV might encounter during its service life is beyond what can be tested, a starting point would involve scenarios centered on core behavioral competencies appropriate for a given ODD. Some interviewees favored this concept; as one put it, “We all take the same driver’s test. Similar [testing] could be done with AVs.” Others noted that AVs, having passed such a test, could have operating restrictions akin to a graduated driver’s license and that the restrictions on new drivers (such as not driving at night or on the highway) are not dissimilar to the constraints associated with ODDs.

But as a functional benchmark, problems with driving tests develop.

The analogy is that when a human has a driving test, the rater is in the vehicle, not outside. The human is being monitored on the process (e.g., do you look over your shoulder when changing lanes). Even if they don’t look, they may not hit something, but the process was not correct [and the rater will notice that and] mark it down as a mistake.

Because AVs are a “black box,” testers have less insight; they can observe only the outcomes during testing, not the process. “Gameability” and cheating are concerns, as is ensuring that the test activities and requirements perfectly match ODD specifications. And frequent updating of

the ADS might raise questions about the frequency of retests. Should “all OTA [over-the-air] updates . . . pass [a test]?” One interviewee argued that “if [the update was] fixing a safety issue, [then] most definitely we should be requalifying [the ADS].”¹⁰⁵ Also, drivers’ tests are historically oriented toward teenage drivers, not the safest group and a far cry from the safe human driver discussed in the previous section.

We are willing to accept 16-year-old drivers, who will learn and get better with experience. This equivalence works less well for AVs “learning.”

Because the number one killer of teens is traffic crashes,¹⁰⁶ the test might not be able to show performance to an acceptably high level.

Finally, validity is questionable. The degree to which passing or failing a driving test reflects overall driving ability is unclear. Although a human-driving test reflects ability to execute key driving maneuvers (behavioral competencies) and a knowledge of driving law, it does not reflect overarching driving skill in every situation.

Despite these issues, as one interviewee noted, at least some of

the industry is pushing for the black-box test [involving some form of scenario testing]. Come up with scenarios that represent an ODD, we’ll test to those, we’ll measure these metrics (safety envelope, Delta-V, etc.) . . . It’s easy to equate it to a human-driver test. You can see, you know the rules, you follow the procedures. Manufacturers are doing all those, and that’s probably where we’ll end up. It’ll be documented that they have done that type of thing. It is unlikely there will be a [public or common] test because everyone’s system is different.

For such testing to be meaningful, the aforementioned issues require further development in terms of scenarios, passing criteria, ensuring nongameability, etc.

The Safety Performance Potential of ADS Technology

One alternative to a threshold rooted in human drivers is a threshold rooted in the performance potential of AV technology—is it as safe as it can possibly be? Such an approach would be subjective and, as one interviewee put it, “define safety of an AV as a thing in itself.” This positions AVs not as the next stage of development of the automobile but as a new thing entirely, to be judged on its own merits. This approach draws from the concepts of “as low as reasonably achievable” (ALARA) or “as low as reasonably practicable” (ALARP)¹⁰⁷ and from the technology performance thresholds set forth for human-driven vehicles in FMVSS.

¹⁰⁵ Establishing a protocol to parameterize the change that an update will make in driving behavior and ensuring that the change will not degrade established safety expectations of the customer or other road users would be a valuable contribution to AV safety.

¹⁰⁶ Centers for Disease Control and Prevention, “Teen Drivers,” webpage, October 28, 2019.

¹⁰⁷ These concepts are covered in detail in Box 5.1 at the end of this chapter.

How can one establish a safety threshold based only on AVs themselves if the technology is still in development and highly varied? On a conceptual level, there are a few options. As one interviewee said:

How safe is “safe enough” is where things get interesting. The minimalist approach is meeting appropriate regulations for a product. Beyond that, it becomes a qualitative argument that depends on the product, what ODD, what features [the developer is] trying to deliver to the market, and the risk assessment that went on.

The qualitative basis for saying that an AV is as safe as it can possibly be could involve evidence that the AV meets all existing best practices, including a variety of technical standards and other best-practice protocols (see Chapter 4). As one interviewee stated, “with regards to outcome, [I] don’t care what steps [the AV] went through but that it was a robust process.” Thoroughness or robustness of process (as described, for example, in a safety case) is certainly one way to consider whether an AV is as safe as it has the potential to be. But there is no mechanism to assess safety-case thoroughness or robustness. At this time, as one interviewee explained, it remains

a judgment call. If you say a vehicle today is “safe,” that means “not a guarantee of no failure.” [It means that] evidence that the rate of failure is low enough that you have trust or confidence in the system. This is very subjective. Assertion is always a judgment call. [But] how does that judgment get made, and by whom?

A threshold for acceptable safety based on the AV technology achieving its maximum potential for safety is theoretically usable. But there is no clear protocol to follow, no established evidence to offer, and no clear guidelines on who should provide the evidence (e.g., a developer or an external evaluator) or to whom it would be provided (e.g., regulators or the general public). Additionally, results from the ALP survey (discussed in Chapter 6 and Appendix A) indicate that thresholds based on technology performance overall (i.e., meeting federal vehicle requirements) were not among the most resonant with the general public.

Safety as Achieving a Threshold Based on an Absolute Safety Goal

Rather than being rooted in the relative threshold of human-driver performance or in AV technology, a threshold could be rooted in an absolute, societally determined goal. Such an approach is essentially one in which the risk falls below an arbitrarily defined probability or burden.¹⁰⁸ Functionally, an absolute goal has utility. Because “we don’t have a standard that translates ‘human drivers’ to AV performance,” rooting instead in an absolute number seems logical. Additionally, such a goal is objective in that the truth of a claim that AVs are safe would not depend on the claimant’s belief (which can be mistaken).¹⁰⁹

¹⁰⁸ Fewtrell and Bartram, 2001.

¹⁰⁹ Möller, Hansson, and Peterson, 2006.

An example that came up frequently in the project interviews is Vision Zero, the aim for zero road traffic deaths by 2050 introduced in Chapter 1. Proponents made such comments as “[our] baseline . . . is zero deaths; few severe injuries. . . . We want actions that are proven to eliminate fatalities and reduce injuries. Vision Zero metrics are fatalities and severe injuries,” and “[our goal is] getting rid of all crashes—I don’t want us to ever be in the position to say, ‘Do this; there will be loss of life, but that’s okay.’”

But having a requirement of zero anything is conceptually fraught and might be practically infeasible.¹¹⁰ Accordingly, interviews with people concerned about the interpretation of a zero-death goal featured such comments as:

The absolute threshold should be zero, though it is not achievable.

[An extension of no deaths is] that “AVs will be perfect” and have zero issues. The problem is that this is an impossible point to meet. Depends how utilitarian or [focused on] expected perfection you are.

No one will wait until things are perfect. They will reach “good enough” and then roll out. Then you must continue monitoring as the process evolves.

I don’t think you’re ever satisfied with safety. If there are any crashes whatsoever, there is still room for improvement.

As long as there are crashes or fatalities, things could be safer.

These comments are not that far from those of people advocating Vision Zero.

Additionally, driving (human and machine) is based on monitoring and assumptions.¹¹¹ For example, it is assumed that the driver in front of us will not suddenly brake at 70 mph on a highway in light traffic, so we do not drive so slowly that we could stop in time, should it happen. It is assumed a pedestrian might illegally cross just as the traffic light is turning green, so we pause before pressing the gas pedal. It is assumed that the vehicle behind us will maintain

¹¹⁰ There might be political reasons for discourse focused on zero risk, as noted by Viscusi: “We are usually assured that our food is ‘safe’ rather than being told that there is a low but nonzero probability that the food will make us ill” (W. Kip Viscusi, *Regulation of Health, Safety, and Environmental Risks*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 11934, 2006, p. 13).

¹¹¹ One potential area for collaboration and convergence between developers and local and state government is in assumptions. As one interviewee stated,

It would be great if [developers] made assumptions based on a standard everyone agreed on. If the assumptions prove wrong, we could revise the standard. We need a standard idea of what is okay to assume about how other actors act.

For example, AV developers could collaborate with local government on environmental factors, such as actual average traffic speed relative to speed limit, all-vehicle following distance of vehicles.

a safe following distance, so we concentrate on what is next to and in front of us. Assumptions useful for ADS operation, explained one interviewee, have a variety of consequences:

Being able to make these assumptions can make AVs flow with traffic well and seem very natural, but it creates a risk that there will be a crash when assumptions are broken. . . . “Safe enough” is about our ability to make assumptions about both AVs and humans. You assume that other drivers will not swerve into you on the road. Some roads are right next to a sidewalk, and you may be going 40 to 45 mph. You assume those pedestrians will not enter the roadway, etc. The parameters for safe driving are making assumptions about these behaviors. For example, when following, you assume that maximum braking is at a certain rate—this affects the follow distance.

If AVs make assumptions while driving, then they will make mistakes, and safety implies that ADS mistakes will seldom happen. This is engineering territory, as one interviewee explained: “If you’re going to have a failure rate, you need to have a plan to mitigate it—duplicate systems or other ways to make it a safe process.”¹¹² But being allowed to make mistakes means that zero anything—issues, errors, hard braking, crash, or death—might be an untenable goal.

RAND has published a national strategic plan based on multistakeholder engagement of what would be involved in achieving Vision Zero by the middle of the century¹¹³—a point when a large AV presence is anticipated. The idea that connects that research to this project is that a goal of zero road traffic deaths is not unattainable; rather, zero traffic deaths and serious injuries are the endpoint of the spectrum of safety. If a threshold is a point that must be exceeded for a condition to be met, Vision Zero is more accurately an ideal rather than a threshold. Although most would agree that a standard of zero traffic deaths and serious injuries is highly desirable, the question remains whether such a high standard is appropriate for AVs, at least in their earlier phases of development. Zero lacks any kind of forgiveness and understanding of the kind that could be factored into metrics (Chapter 3) and processes (Chapter 4). Applying it to AVs prematurely might incentivize companies either to ignore the threshold or to put forward a product so conservative in conduct that no one uses it. Additionally, no country or region has achieved zero road traffic deaths to date. That said, much benefit has and will come from setting Vision Zero as a goal to direct behavior of multiple stakeholders toward the continual improvement of safety for the overall transportation system.

If zero is not a plausible threshold for AV safety, even as it continues to act as a societal goal, what other options are there? One possibility is the concept of GAMAB, or *Globalement Au*

¹¹² Assumptions feature in technical standards-setting. Jack Weast, who chairs the IEEE P2846 Working Group mentioned in Chapter 4, developed a presentation for the Automated Vehicle Symposium 2020 that relates metrics to assumptions (Jack Weast, “Metrics and Assumptions in Safety Assurance,” presentation, July 29, 2020).

¹¹³ Ecola et al., 2018.

Moins Aussi Bon (French for “generally at least as good as”).¹¹⁴ As discussed in one interview, GAMAB requires the object of inquiry to be as “safe as comparable technology.” In other words, the new technology or system must include the same risk as the old technology or system.¹¹⁵ The theory is that one cannot get below this level. A hybrid between a threshold rooted in the average human driver and one rooted in the absolute goal of the existing crash fatality count, GAMAB might have the conceptual negatives of both without either’s positives—it is not highly communicable or evocative, and it tacitly accepts the existing level of death from motor vehicle crashes.

An alternative is minimum endogenous mortality (MEM), which, like GAMAB, is “helpful because you have a value.” Although GAMAB is rooted in socially accepted risk, MEM focuses on all-cause death rates.¹¹⁶ Parsing MEM, *endogenous mortality* refers to age-specific, all-cause national death rates, and “minimum” refers to when this rate is at its lowest. As applied, MEM also factors in known contextual risks.¹¹⁷ As explained by an interviewee, under MEM,

[a] new system [e.g., a semiautomatic system in trains] should have a value of MEM/20 [which factors in 20 contextual risks]. . . . The theory is that you cannot get below this mortality rate. . . . The common way this is used is that any new technology cannot lower the minimum endogenous mortality. . . . Any new tech should not increase MEM or not increase it by x percent [in this case, 1/20].

In other words,¹¹⁸ the mortality threshold is rooted in demography. As this quotation illustrates, MEM has been used elsewhere in transportation. Along with GAMAB and ALARP (Box 5.1), it is an option presented in UL4600.

A virtue of MEM and GAMAB is their recognition that life is full of risk. Taken to an uncomfortable extreme, MEM defines “how many people are you allowed to kill.” Overall, there is an ethical imperative to reduce the number of road traffic deaths by reducing risk and not causing new risk. If one sets a nonzero safety threshold specifically rooted in fatalities, does it tacitly acknowledge that society finds that number of deaths from car crashes acceptable? If the accepted amount is around 37,000 annually, then can AVs be involved in 20,000 crash deaths? Or 1,000? Where would society draw the line in terms of a tolerable number of deaths? Or is any reduction, regardless of how large, sufficient under the argument that at least one life saved is

¹¹⁴ Philipp Junietz, Udo Steininger, and Hermann Winner, “Macroscopic Safety Requirements for Highly Automated Driving,” *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2673, No. 3, 2019.

¹¹⁵ Junietz, Steininger, and Winner, 2019.

¹¹⁶ Junietz, Steininger, and Winner, 2019.

¹¹⁷ Pegasus, “Requirements & Conditions—Stand 7: Social Acceptance for HAD (L3),” presentation at Pegasus Symposium 2019, University of Glasgow, May 14, 2019.

¹¹⁸ If MEM is, for example, 6×10^{-5} person-years (as it is in Germany) and the goal is MEM/20, then the acceptable fatal crash risk is a risk less than 1/20 of the above (or 3×10^{-6} person-years). Pegasus, 2019.

enough? Several interviewees touched on the politics of the problem, as illustrated by the following quotation:

Industry hype and marketing say that AV will be perfect, with zero mortalities or problems. That is an impossible ideal to meet. No legislator wants to be the one to say that 40 deaths are okay, as opposed to zero, even though original number of human-driven car deaths [is] 40,000. . . . Even 40 sounds like too many to some—it's a political concern. [In contemplating how AVs help with the] elimination of faulty human judgment—how utilitarian should we be or how much should we expect perfection out of machine?

Providing further complication, another argument could be made that as long as AVs do not cause more deaths,¹¹⁹ the other benefits of AVs (low cost, accessibility, convenience, etc.) make them worthwhile.¹²⁰

How Thresholds Can Be Used

Each threshold has strengths and weaknesses. Conceptually, human-driver performance is attractive because human drivers are what we know and serve as the alternative. But as we have discussed, the definition of an average human driver or even a good human driver is disputed and, except for in a few localities, information is insufficient for it to function as a threshold. Conceptually and functionally, ADS technological performance has potential as a threshold, but work is still ongoing to develop best practices. Furthermore, absolute goals provide an easily used and proven threshold, but one that is unlikely to find general acceptance (Table 5.1).

Other types of thresholds exist, such as meeting a product warranty or insurability. Design of a warranty for an AV ride-share service business model could presumably spell out key elements (e.g., transport to a destination in a reasonable amount of time and without harm to the rider). Insurability is complicated by a lack of actuarial data around how to quantify risk. Such mechanisms would be matched to the business model at least as much as to the design and engineering of the AV (which, in turn, will be influenced by business model expectations).

¹¹⁹ Nidhi Kalra and David G. Groves, *The Enemy of Good: Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles*, Santa Monica, Calif.: RAND Corporation, RR-2150-RC, 2017.

¹²⁰ Much of the discussion around absolute safety goals focuses on lagging events: injuries and deaths. However, there is arguably strong utility in focusing on leading events. As already stated, leading events occur much more frequently than lagging events, allowing statistically valid assessments to be made with less exposure (VMT). Additionally, using leading measures with absolute safe goals solves a dilemma: As articulated earlier in this chapter in the section titled “Safety as Achieving a Threshold Based on Human Driving Performance,” one of the main challenges of using leading events is that comparable human-driver data are rarely available. Using leading events in concert with absolute safety goals instead of thresholds based on human-driving behavior nullifies this challenge.

Table 5.1. Ability of Each Threshold Type to Provide Evidence of Acceptable Safety

Threshold	Conceptually	Functionally
Human drivers	Strong	Weak
ADS technology performance	In development	In development
Absolute goals	Weak	Strong

And yet, a threshold is valuable for assessing safety. As one interviewee stated, “To put a number or letter on it [a threshold] has always been a bit artificial, even though that is what people are looking for.” In the absence of good alternatives, there are four questions to ask in determining how a threshold can provide an approach to AV safety:

1. What role do absolute goals play?
2. How can thresholds rooted in ADS technological potential and in human drivers work together?
3. Should every developer use the same thresholds?
4. Regardless of the type of threshold, is everything that does not meet the threshold of acceptably safe automatically unsafe?

Regarding the first question, the role of absolute goals varies depending on what the aim is. If the aim is zero, it serves as an ideal. It is, in the words of the National Safety Council, a “moonshot goal” and a rallying point for transportation safety stakeholders across the board.¹²¹ Although zero is not usable as a threshold for AVs, which are only one component of the complex transportation ecosystem, “AVs [can serve] as a tool to help achieve Vision Zero’s goal of zero deaths.”

The second question is part of a larger discussion of how measurement and process approaches from Chapters 3 and 4 work together. This interplay and mutual support is discussed in detail in Chapter 7.

The third question is fraught, given the AV industry’s diversity and given the stakes, as forecast by one interviewee: “any metric will lead to a PR [public relations] war between AV developers to come out with a metric that is infinitesimally superior.” Interviewees from all stakeholder categories called for a blend of uniformity and individuality.¹²² The following quotation illustrates this:

Every company would like some kind of flexibility with regards to the argumentation they make, but variations on very similar content. If [one gives questions] to ten engineers well versed in state of the art, the answers may be similar.

¹²¹ National Safety Council, “Happy Anniversary, Road to Zero!” *Safety First*, blog post, October 23, 2017.

¹²² This tension is noted Fraade-Blanar et al., 2018.

As these and other comments indicate, flexibility is essential for human-driver and ADS-technology potential thresholds to be accepted for broad use. After all, “if someone has a creative way of demonstrating safety, we should let them” use it.

Definitions, by contrast, might benefit from uniformity: “To a certain extent, all need to be speaking the same language. And the language is important to not misrepresent to people what the vehicles do.” Because companies will have such diverse ODDs, use cases, business models, technology, etc., the way in which evidence for meeting safety thresholds is shown and discussed should be uniform and transparent for all audiences.¹²³ Having uniform definitions will assist in this effort.

The fourth question is philosophical and relates back to the original conception of safety. The idea of being acceptably safe has an unspoken component: safe enough for a given activity, which might be any number of things (e.g., commercial deployment, driving on public roads, removing a safety driver, extending to a new ODD, expanding the number of vehicles in services). A system that does not meet thresholds of safety might be safe for other activities but not safe enough for X. Consequently, it is possible that different thresholds work best for different activities and vary depending on the nature of the activity and who (e.g., policymakers, the public, a company itself) needs evidence of safety. This situation requires a stepped approach:

All the manufacturers have objective measures of safety in house. . . . It may be enough to prove that the system is safe enough for the next round of testing, but maybe not enough to put 100,000 new level 4s on the road unsupervised. I’m distracted by what happens down the road with mixed modality of both humans and machines on the road—lots of variety of human drivers, software drivers, and environmental conditions. A stepped approach would be helpful.

Developers today take a stepped approach that is based on their assessment of safety evidence.

As already discussed, the original view of a safety threshold is that it evolves. Part of this evolution will reflect advances in technology; another part of it will reflect expansions and advancement of services and the business of AVs, and yet another part will reflect changes in the behavior of human drivers sharing the roadway with AVs. Thresholds of safety, then, reflect what is acceptably safe for the existing type and level of operation (e.g., safe enough for now) and/or acceptably safe for an expansion or removal of safeguard (e.g., safe enough for the next stage).

¹²³ This is already a challenge in the AV industry. Eric R. Teoh, “What’s in a Name? Drivers’ Perceptions of the Use of Five SAE Level 2 Driving Automation Systems,” *Journal of Safety Research*, Vol. 72, 2020.

Box 5.1. ALARA and ALARP

When we move the discussion beyond specific standards, the concept of a threshold rooted in technology potential could borrow from another industry, such as ALARA or ALARP. These concepts have been used in such areas as nuclear power and toxic contaminants, and they focus on what is doable given the sophistication of available technology, economic pressures, and what the public will accept.^a ALARP, especially, tries to elucidate a level of risk between unacceptably high and acceptably low.^b The natural corollary is that the highest risks are entitled to the most resources aimed at risk reduction.^c

Although the literature holds examples of assignable quantitative values (e.g., somewhere between the existing risk of motor vehicle fatality and the risk of being struck by lightning),^d most uses of these concepts are qualitative, focus on standard practices and processes, and are area- and industry-specific. In this, ALARP and ALARA differ from GAMAB and MEM, which are referenced together in UL4600 and other technical literature.^e GAMAB and MEM are specifically quantitative and thus easier to apply to AVs and easier to identify whether AVs meet (or fail to meet) their threshold.

ALARP might be difficult to achieve for new technology or for developers to be able to identify when they have achieved it.^f It did not resonate broadly with the people we interviewed. Most were unfamiliar with the concept; among those who had heard of it, enthusiasm was minimal. As interviewees said:

There is a lot in the word “reasonably.” What is practical? What is reasonable? “Practical” means you met the point of diminishing returns.”

[The] problem is the stuff you didn’t think of, incompleteness of hazard analysis. . . . [Y]ou might miss a hazard, it happens all the time.

How is “reasonable” determined? What does “practical” mean? It is a fine qualitative standard to hold up to, but difficult to argue in a court of law.

The practicalities of ALARA and ALARP are not exactly pertinent for ADS. It makes sense for radiation exposure or something, but it doesn’t translate over as well for driving systems.

That’s how we operate, but it can lead to disasters . . . like with nuclear power plants in Japan. Cleaner energy than fossil fuels, but still risky.

However, there is utility in these concepts. Safety case architecture, such as UL4600 and PAS 1881, include references to ALARP as a risk evaluation option or goal.^g As one interviewee noted, the “positive trust argument says that we’ll never have enough data in the real world to ‘prove’ it is safer than human-driven vehicles. So, part one of positive trust argument is ‘be as safe as you can be.’” Arguably, the utility of the ALARP and ALARA concepts lies in this positive trust argument; the “as low as” rather than the “reasonably practical” or “reasonably acceptable,” which, for AVs, sits on the spectrum between undefined and unhelpfully emotive and imply moral trade-offs between benefits and costs.^h A version of ALARP and ALARA for AVs would:

- tolerate something higher than zero risk but less than a given upper level
- provide guidance as to level of risk, albeit qualitative in nature, somewhere in between the two
- understand that some uncertainties must be accepted
- expect that resources will go to addressing the highest risks.

With this list in mind, phrasing an approach along the same lines as ALARA and ALARP might have utility in enhancing understanding of that approach and in connecting it to approaches already in use, albeit in other industries (such as nuclear energy and toxicology). Table 5.2 provides translations of each of the approaches into possible variations of ALARA and ALARP.

^a Philipp Matthias Junietz, *Microscopic and Macroscopic Risk Metrics for the Safety Validation of Automated Driving*, doctoral thesis, Darmstadt: Technische Universität, 2019; Robert E. Melchers, “On the ALARP Approach to Risk Management,” *Reliability Engineering & System Safety*, Vol. 71, No. 2, February 2001.

^b Terje Aven, “On the Ethical Justification for the Use of Risk Acceptance Criteria,” *Risk Analysis*, Vol. 27, No. 2, April 2007; Hendrik Schäbe, *Different Principles Used for Determination of Tolerable Hazard Rates*, Cologne, Germany: Institute for Software, Electronics, Railroad Technology, 2001.

^c Aven, 2007.

^d Junietz, Steinger, and Winner, 2019.

^e Underwriters Laboratories Standards, 2020; Andrew J. Rae, *Acceptable Residual Risk—Principles, Philosophies and Practicalities*, paper presented at second Institution of Engineering and Technology International Conference on System Safety, London, 2007; Junietz, 2019.

^f Melchers, 2001.

^g Underwriters Laboratories, “Presenting the Standard for Safety for the Evaluation of Autonomous Vehicles and Other Products,” webpage, undated-b; British Standards Institution, *Assuring the Safety of Automated Vehicle Trials and Testing—Specification, PAS 1881*, London, 2020. This specification establishes minimum requirements for AV testing safety cases.

^h Melchers, 2001.

Table 5.2. Applying ALARA and ALARP Phraseology to Approach

Existing Approach		Approach as a Risk Level
Safety as a measurement	→	As safe as calculatable quantitatively OR As quantifiably safe as demonstrable
Safety as a lagging measure	→	As safe as calculatable quantitatively by a lagging measure OR As quantifiably safe as demonstrable by a lagging measure
Safety as a leading measure	→	As safe as calculatable quantitatively by a roadmanship measure OR As quantifiably safe as demonstrable by a roadmanship measure
Safety established and/or endorsed by a process	→	As comprehensively safe as a process can establish OR As safe as evidence of a process can support
Safety established and/or endorsed by technical standards	→	As safe as evidence of a process can support using technical standards OR As safe as showing evidence of a technical standards process can support
Safety established and/or endorsed by government regulation	→	As safe as legally required
Safety established and/or endorsed by safety culture	→	As believably safe as the company's safety procedures demonstrate OR OR As safe as a safety culture can confirm OR As safe as a company can be
Safety as achieving a threshold	→	As safe or safer than the threshold prescribes OR As safe as acceptable
Safety as achieving a threshold predicated on human driving	→	As safe as the average or safe human driver
Safety as achieving a threshold predicated on technology	→	As safe as the technology can be

6. Communicating About Safety

In humans, the classification of an object or action as being safe is inextricably linked to the perception of risk. To the greater public, a sense of safety often is not a judgment made after a systematic risk assessment but one that is developed in response to emotional and affective cues.¹²⁴ Particularly in the realm of new technologies, where individuals lack prior information and experience, these affective responses are sensitive to communication from government, experts, media, and industry stakeholders. The content and source of communication pertaining to safety both play an integral role in the public perception of safety as meeting an acceptable standard.

One stakeholder expanded on this topic and noted that “safe enough” is, at its heart, a subjective statement. To paraphrase: Measurements provide us values that give us an understanding of the probability of a hazard, but the determination that such a probability is “safe” relies on human judgment. A declaration that something is safe is not a guarantee that there will not be a failure; rather, it is a statement that the rate of failure is low enough that there can be trust or confidence in the system. How does that judgment get made, and by whom? Where were data acquired, and what kind of argument was made to indicate that the safety assertion was adequate?

Risk Perception

Risk itself is a psychological construct, not an objective reality. As explained by Paul Slovic, a leading researcher in the area of risk perception:

Risk does not exist “out there,” independent of our minds and cultures, waiting to be measured. Human beings have invented the concept *risk* to help them understand and cope with the dangers and uncertainties of life.¹²⁵

Cognitive psychologists and behavioral economists have uncovered several heuristics and biases that affect our perception of risk outside an understanding of the rate of negative outcomes. The study of risk perception has been conducted largely in the fields of cognitive psychology and behavioral economics, and it has produced conceptual theories and models of risk, which we discuss briefly in the ensuing sections.

¹²⁴ Paul Slovic, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor, “Risk as Analysis and Risk as Feelings: Some Thoughts About Affect, Reason, Risk, and Rationality,” *Risk Analysis*, Vol. 24, No. 2, 2004.

¹²⁵ Paul Slovic, “Perceptions of Risk: Reflections on the Psychometric Paradigm,” in Sheldon Krinsky and Dominic Golding, eds., *Social Theories of Risk*, New York: Praeger, 1990, p. 121.

Psychometric Paradigm of Risk

The question of “how safe is safe enough?” has been thoroughly investigated in the scientific literature, largely prompted by a 1969 article investigating trade-offs between the benefit and risk provided by new technologies.¹²⁶ As Paul Slovic summarized,

Starr concluded that (a) acceptability of risk from an activity is roughly proportional to the third power [cube] of the benefits from that activity; (b) the public will accept risks from voluntary activities (such as skiing) that are roughly 1,000 times as great as it would tolerate from involuntary activities (such as food preservatives) that provide the same levels of benefits; and (c) the acceptable level of risk is inversely related to the number of persons exposed to the risk.¹²⁷

Researchers later expanded this to include other dimensions found to influence risk perception. Along with voluntary or involuntary exposure, aspects concerning the immediacy of the effect, whether the risk is known by both the individual and to science, dread of the outcome, number of people experiencing the outcome concurrently, and the severity of consequences were all combined to create a theoretical approach to the study of risk perception referred to as the *psychometric paradigm*.¹²⁸

These dimensions capture the idea that what people know and how they feel about risky activities, the potential outcomes, and the ability to avoid negative outcomes all directly affect their perception of risk. In many ways, AVs and human-driven vehicles carry the same risks—a vehicle accident is still a vehicle accident regardless of who is driving—but they differ in one key area. AV technology is still in its relative infancy, and little is known about precise rates of adverse outcomes. When the likelihood of adverse events is unknown to individuals and science, the perception of risk increases. Although human-driven vehicles are still inherently risky, the sense people have that they know what to expect in relation to adverse events (e.g., crashes) lowers their perception of risk associated with driving. Over time, as more data are collected about the likelihood of AV accidents and their crash rates become known to the public, perceptions of risk for AV travel should fall and even surpass that of human-driven vehicle travel (as the safety record is shown to surpass that of human drivers).

Affect Heuristic

Building on the concept of the psychometric paradigm, the affect heuristic further explores and explains the relationship between emotions and the associated risks and benefits associated

¹²⁶ Chauncey Starr, “Social Benefit Versus Technological Risk,” *Science*, Vol. 165, No. 3899, 1969.

¹²⁷ Slovic, 1990, p. 3.

¹²⁸ Baruch Fischhoff, Paul Slovic, Sarah Lichtenstein, Stephen Read, and Barbara Combs, “How Safe Is Safe Enough? A Psychometric Study of Attitudes Towards Technological Risks and Benefits,” *Policy Sciences*, Vol. 9, No. 2, 1978.

with an activity.¹²⁹ People’s knowledge about the risks associated with an activity or technology affects their emotions (e.g., risky things prompt negative feelings, safe things prompt positive feelings), which in turn colors their perception of the benefits provided by the risky activity or technology. This also works in the opposite direction. In a demonstration of this effect, participants in one study were given information pertaining to the risks and benefits associated with a technology.¹³⁰ When the risk was described to be low, the participants rated the associated benefits as high, and when the risk was described to be high, the associated benefits were perceived to be low. Likewise, in the opposite direction, describing the technology as being highly beneficial yielded responses associated with lower perceptions of risk; descriptions stating that the benefits of the technology were low yielded responses associated with perceptions that risk was high.

This heuristic applies to AVs in a way that differs from many of the other approaches discussed in this report to establish and communicate safety. Many of those approaches focus on establishing a method of quantifying or signaling that AVs are safe—or, in other words, low-risk. This heuristic shows that in the case of communicating safety to the public, establishing the benefits provided by AVs might serve as a method of addressing risk perception that is an alternative to focusing efforts on communicating the rates of negative outcomes.

Individuals’ baseline expectations of the benefits provided by AVs will also affect perceptions of risk. Subsets of the public with restricted access to transportation, particularly those who are unable to drive, should see AVs as providing greater potential benefits compared with those groups that have relatively unrestricted access. As a result, individuals in the former group (e.g., older adults, vision-impaired people, unlicensed adults, adolescents) should have a more positive attitude toward AVs and perceive the associated risks to be lower. This effect should also be present for temporary driving restrictions (e.g., choosing between driving or taking an AV when inebriated, temporarily losing access to a normally available vehicle), when the circumstances will raise perceptions of benefits, increase positive attitudes, and lower perceptions of risks associated with AVs.

Cognition and Safety Statements

Because humans are not rational decisionmakers, biases and heuristics play an important role in evaluating novel scenarios and decisions under risk. Decisionmaking heuristics—such as loss aversion, anchoring, and availability—will all play large roles in the campaign to communicate to the public about AV safety. Framing messages in positive or negative language independently

¹²⁹ For example, see Slovic et al., 2004.

¹³⁰ Melissa L. Finucane, Ali Alhakami, Paul Slovic, and Stephen M. Johnson, “The Affect Heuristic in Judgments of Risks and Benefits,” *Journal of Behavioral Decision Making*, Vol. 13, No. 1, 2000.

influences decisionmaking about risk.¹³¹ Traditionally (as touched on in Chapter 3), people communicate safety messages in terms of negative incidents that were avoided (Safety I) rather than the safety capability of a system in a complex and fluid environment (Safety II).

The perception of control is generally associated with perceived lower risk,¹³² and it is understood to play a large role in the public's willingness to accept relatively high levels of risk when driving.¹³³ There is some evidence that perceived control has two separate aspects. Nordgren and colleagues named these "volition" (influence over exposure to the risk itself, which is associated with an increase in perceived risk) and "control" (influence over the outcome of the risky event, which is associated with a decrease in perceived risk).¹³⁴ Using this terminology, passengers in any vehicle likely have volition (e.g., to ride or not) but do not have control over the outcome. In this sense, riding in an AV should be no different from riding in a vehicle driven by another human being. Trust of the individual or system will likely play a larger role in the perception of risk than in the perception of being unable to control the outcome.

In general, numerical and mathematical abilities vary considerably across all demographics. Widespread adoption of AVs will necessitate communication of safety and risk that accounts for these variations and must reach individuals whose ability to interpret such information varies considerably. Conventional motor vehicles have addressed this issue with definitive statements from trusted authoritative bodies (e.g., meeting FMVSS) or the use of simplified, culturally relative rating scales (e.g., five-star, letter grade). To be useful, the sources of these statements must be seen as credible and trusted, a status that is not easily cultivated.

Other industries that operate within a spectrum of acceptable risk seem to have bypassed the need to convince the public about their location on a safety spectrum. Nonautomotive industries (e.g., amusement parks, pharmaceuticals) do not seem to get the same scrutiny from the public. The presence of the product in the market is seen by the public to imply that a safety threshold has been met and that the product is safe to use. As one interviewee remarked, "[A]viation doesn't compete on safety. People don't get on an airplane looking for the 5 stars on the side of the cockpit—they don't think that way in aviation." Another commented on amusement parks: "If you go to an amusement park and get on a roller coaster, you assume that someone has checked that it is safe enough." In the case of aviation and amusement parks, it is possible that the business design (e.g., paying for rides on a machine owned by a corporation and regulated by

¹³¹ Amos Tversky and Daniel Kahneman, "The Framing of Decisions and the Psychology of Choice," *Science*, Vol. 211, No. 4481, 1981.

¹³² For example, see Starr, 1969; and Neil D. Weinstein, "Why It Won't Happen to Me: Perceptions of Risk Factors and Susceptibility," *Health Psychology*, Vol. 3, No. 5, 1984.

¹³³ Ian Duncan, "EasyMile Autonomous Shuttles Barred from Carrying Passengers," *Washington Post*, February 29, 2020.

¹³⁴ Loran F. Nordgren, Joop van der Plight, and Frenk van Harreveld, "Unpacking Perceived Control in Risk Perception: The Mediating Role of Anticipated Regret," *Journal of Behavioral Decision Making*, Vol. 20, No. 5, 2007.

government) communicates to the public that due diligence has been done on safety by a trusted authoritative body. For AVs, a more direct message of safety to the public might be required for direct-to-consumer sales, although less communication might be necessary for a ride-share model.

Habituation and Its Effect on Risk Perception

Personal experience colors the perception of risk. Acclimation of the public to automation will be a primary goal at the outset of greater adoption. Specific to AVs, the first experience is of particular importance for cultivating trust and acceptance.¹³⁵ Stakeholders overwhelmingly agree that exposure is an important factor in the acceptance of AVs. One noted that, although people report a 70-percent unfavorability rating nationwide, surveys querying people who live in areas with AV deployment report 70-percent favorability. Another reported that there is

[a] spectrum of consumer interest, from “I’ll never get in” to “can I have one tomorrow?” When you get into the car, you’re nervous, but then after five minutes you’re bored. After ten minutes, you almost forget you’re in an AV.

Of course, these initial experiences must be positive to increase acceptability. One stakeholder mentioned a test of autopilot technology in which the vehicle drove toward a stop sign with children present. “In this case, exposure showed real flaws.” The cost of negative first impressions can be dire. Along with anecdotal accounts of the role of first impressions in business, the psychological literature has also historically found that the first information received is disproportionately persuasive compared with subsequent information and that first impressions are highly predictive of final assessments.¹³⁶

Attitudes Toward Technology

Sustained use of a technology is preceded by the perception of the technology and the decision to adopt the technology. Communication of safety is of particular importance in the promotion of technology acceptance and adoption in areas in which negative outcomes, however rare, have the potential for severe harm. Acceptance of technology can be negatively affected by a negative reputation (of the technology or its provider) or a conceptual link between the technology and negative outcomes, regardless of the actual rate of harmful events. For example, in the 1970s, the acceptance of microwave ovens was stymied when they received a reputation of

¹³⁵ Franziska Hartwich, Claudia Witzlack, Matthias Beggato, and Josef F. Krems, “The First Impression Counts—A Combined Driving Simulator and Test Track Study on the Development of Trust and Acceptance of Highly Automated Driving,” *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 65, August 2019.

¹³⁶ Frederick Hansen Lund, “The Psychology of Belief: A Study of Its Emotional and Volitional Determinants,” *Journal of Abnormal and Social Psychology*, Vol. 20, No. 1, 1925; Vernon A. Stone, “A Primacy Effect in Decision-Making by Jurors,” *Journal of Communication*, Vol. 19, No. 3, 1969.

causing harm after media response to a Consumer Reports recommendation that they be avoided because of safety concerns.¹³⁷

Contemporary Models of Technology Acceptance

Along with perceived benefits, perception of associated costs is an integral predictor in contemporary models of technology acceptance (e.g., the updated version of the Unified Theory of Acceptance and Use of Technology [UTAUT2]¹³⁸). Though these constructs were originally intended to capture the performance benefits and the ease of use of a novel technology, researchers investigating the acceptance of mobile banking technologies and ADAS soon extended the model to include variables measuring the perception of risk and trust. In these models, risk perception is included as an individual predictor, separate from perceived costs, and explains a unique portion of the variance in the intention to use technology and subsequent usage behavior.

Specific to automation, the technology acceptance model (TAM)¹³⁹—a conceptual predecessor of UTAUT2—was extended to specifically address automation in the automation acceptance model.¹⁴⁰ This model expanded on the original by including variables that capture trust of the automation and its compatibility with the task at hand. In this model, compatibility can be hindered both by exceeding the operator's needs for autonomy and by falling short of those needs. Communication to end users that increases a sense of trust and provides information about the compatibility of specific levels of AV automation should increase the acceptance of AV technology. Additionally, this model includes feedback from actual system use. Along with positively affecting perceptions of the ease of use and usefulness of automation, this model explains that system use also increases trust and the perception of compatibility (as users develop a greater understanding of the abilities of the system).

Communication from the Automated Technology Itself

Communication of the automated capability of vehicles is an important way to address perceptions of risk, particularly in a mixed automation and human-driven traffic environment. Communication can be achieved through both the passive appearance of the vehicle (e.g., visible sensors) and the active communication of performance or intention from automated systems, and

¹³⁷ John M. Osepchuk, "The History of the Microwave Oven: A Critical Review," *2009 IEEE MMT-S International Microwave Symposium Digest*, 2009.

¹³⁸ Viswanath Venkatesh, James Y. L. Thong, and Xin Xu, "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology," *MIS Quarterly*, Vol. 36, No. 1, 2012.

¹³⁹ Fred D. Davis, Richard P. Bagozzi, and Paul R. Warshaw, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Management Science*, Vol. 35, No. 8, 1989.

¹⁴⁰ Mahtab Ghazizadeh, John D. Lee, and Linda Ng Boyle, "Extending the Technology Acceptance Model to Assess Automation," *Cognition, Technology & Work*, Vol. 14, No. 1, 2012.

preferred methods depend on the intended audience. Pedestrians show preference for overt visual communication, both in active electronic messaging that a vehicle is piloted by automation (e.g., an electronic human-machine interface [eHMI]—such as a conspicuous light bar that glows in a different pattern to communicate that the vehicle is yielding or not yielding) and, to a lesser extent, in conspicuous sensor systems.¹⁴¹ Inside the vehicle, where signs of automation are less conspicuous in interior design (at least until the removal of the steering wheel in fully automated level 5), communication of intent of the ADS is associated with increased ratings of trust and preference but does not decrease anxiety or level of driving effort in novice users.¹⁴² The accuracy of messages is also of particular importance because inaccurate communication can reduce user trust. Although absolute accuracy is not likely to ever be maintained, it is important to note that designers should err on the side of caution—misses are more harmful to trust than are false alarms.¹⁴³

Consumer Understanding of Existing and Prospective Advanced Driver Assistance System Capability

Consumers have begun to experience automation in some aspects of driving through ADAS in newer vehicles. Although ADAS presume the engagement of a human driver, these steadily improving features expose people to what happens when such functions as lane-keeping or emergency braking come under computer control. As consumer products that are freely available for purchase without additional constraints or licensure, vehicles equipped with ADAS that were available in 2020 often reached consumers without specific training on the appropriate use or capabilities of the product or system. A recent survey found that only about one-half of new car buyers were offered system-specific training for their new vehicles by the dealership.¹⁴⁴ In the same report, the greater proportion of system users indicated that they either did not seek out information on the system operation or learned to use it through trial and error. For these users, the name of the system might be the only formal communication they receive about its capability.

¹⁴¹ Sander Ackermans, Debargha Dey, Peter Ruijten, Raymond H. Cuijpers, and Bastian Pfleging, “The Effects of Explicit Intention Communication, Conspicuous Sensors, and Pedestrian Attitude in Interactions with Automated Vehicles,” *Proceedings of the CHI 2020 Conference on Human Factors in Computing Systems*, Paper 70, 2020.

¹⁴² Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K. Pradhan, X. Jessie Yang, and Lionel P. Robert Jr., “Look Who’s Talking Now: Implications of AV’s Explanations on Driver’s Trust, AV Preference, Anxiety, and Mental Workload,” *Transportation Research Part C: Emerging Technologies*, Vol. 104, July 2019.

¹⁴³ Herbert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T. Esterwood, X Jessie Yang, Lionel P. Robert Jr., and Dawn M. Tilbury, “Comparing the Effects of False Alarms and Misses on Human’s Trust in (Semi)Autonomous Vehicles,” *HRI ’20 Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, New York: Association for Computing Machinery, 2020.

¹⁴⁴ Ashley McDonald, Cher Carney, and Daniel V. McGehee, *Vehicle Owners’ Experiences with and Reactions to Advanced Driver Assistance Systems*, Washington, D.C.: AAA Foundation for Traffic Safety, September 2018.

The names given to ADAS independently influence the expectations of their capabilities by the general public.¹⁴⁵ Although these names can be used to signal functionality to users, the accuracy of the nomenclature in referring to the capabilities of the system is a key factor for safe use by the general public. For example, recent research has shown that the name “autopilot,” compared with other names, has specifically been associated with an increased likelihood of reporting that several nondriving behaviors are safe to perform during system operation.¹⁴⁶ Two interview sources also directly referred to “autopilot” as being a misleading name for a level 2 system, remarking that “the marketing of current ADAS is misleading. Tesla’s Autopilot is not an actual autopilot” and that one “issue is when a manufacturer releases something they call ‘autopilot’ and aren’t clear about what it can do, that’s dangerous.” Recognizing the importance of reliable and valid nomenclature for advanced driving technologies to the understanding of the public, four major safety and consumer advocacy groups recently agreed on standardized naming conventions for 20 ADAS in five different categories.¹⁴⁷

Convincing the Public

Sources of Communication—ALP Survey

Communication about the safety of AVs comes from a multitude of sources with varying public perceptions of trust and credibility. Types of information are not necessarily connected with their ultimate sources in public understanding (for example, the existence of government safety standards cannot be assumed to be linked in public understanding to NHTSA, nor is it expected that the public would think through how industry messaging might inform both government officials and public awareness). A further obfuscation of the matter is that individuals do not always consciously recognize their own biases.¹⁴⁸

In a survey study conducted with the RAND ALP,¹⁴⁹ data were collected from more than 2,200 participants measuring preferences for different sources of AV safety information and the relative influence of safety messages from each of these sources on the perception of AV safety. The influence of safety messages from different origins was measured implicitly. The questions asked participants to indicate their perceived safety of AVs after receiving various patterns of information from eight different sources: AV crash rates (a lagging measure); AV near-miss rates (a leading measure); federal vehicle requirements; friends and family; and official positions

¹⁴⁵ Hillary Abraham, Bobbie Seppelt, Bruce Mehler, and Bryan Reimer, “What’s in a Name: Vehicle Technology Branding & Consumer Expectations for Automation,” *Proceedings of the 9th ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2016.

¹⁴⁶ Teoh, 2020.

¹⁴⁷ American Automobile Association et al., undated.

¹⁴⁸ For an overview, see Daniel Kahneman, *Thinking, Fast and Slow*, New York: Macmillan, 2011.

¹⁴⁹ Details are provided in Appendix A.

from the federal, government, state and local governments, safety advocacy groups, and AV companies. In the following description of the ALP study and in Appendix A, these eight variables are referred to as “safety message sources.” Although this set of sources includes both primary data sources (e.g., AV crash rates, AV near-miss rates) and secondary sources (e.g., statements from government or AV companies), we treat safety information originating from these sources as equivalent and independent in the statistical models. In the real world, there are or likely will be instances in which some of the secondary sources used in the survey draw on information from one of the primary sources in their messages to the public about the safety of AVs. By separating these variables in the design, we allow for these safety messages to potentially disagree, as they sometimes do in the real world. For example, in municipalities testing AVs, the local governments have communicated to the public that AVs are safe even though no information is available from AV near-miss rates. This topic is further discussed in Appendix A.

To calculate the relative influence of each source, the safety messages and resulting perceived safety ratings were entered into a linear mixed model. Using this method, a regression coefficient signifying the isolated influence of each safety message source on perceived safety of AVs could be calculated and compared with each other. Additionally, preference for the message sources was measured explicitly; participants were asked to rank order their preference for information about AV safety originating from the same eight sources used for the implicit measure. For an in-depth description of the methods and results, see Appendix A.

Overall, participants provided perceived safety ratings for 20 different experimental items, each with a mixture of safety messages (from “strongly show that AVs are not safe” to “strongly show that AVs are safe”) from the eight different sources. Overall, participants’ responses indicated that they were generally skeptical of AV safety. Only one of the 20 items had an average perceived safety score (range: 0–100) that was greater than the midpoint of “generally safe.” This item described a situation in which each of the eight sources provided a message strongly showing that AVs are safe and had an average perceived safety score of 74.9, indicating that participants perceived AVs to be “very safe” on average. The average score for all 20 items was only 31.2, indicating that AVs were on average perceived to be “a little safe.” Further discussion of perceived safety scores for individual items can be found in Appendix A.

Comparing the influence of the message sources using the implicit measures, the messages from all eight sources about whether AVs are safe had a significant positive relationship with perceived AV safety when analyzed using a linear mixed model to account for differences in the overall perception of AV safety between individuals. As the message shifted from strongly negative to strongly positive, each source was associated with increases in perceived AV safety when controlling for variation in the other seven sources. There was variation in the strength of that association, or the ultimate influence, between the sources. Four of the sources had stronger effects; the other four were markedly weaker. The safety message source with the largest influence was AV crash rates. Information from state and local governments had the next highest

influence, followed by information from the federal government and AV near-miss rates, which had similar influences. Information from AV companies had the least influence on perceptions of safety.

We also asked participants to rank order the sources depending on their preferred source of information on AV safety. When we measured the responses explicitly, we found an interesting pattern of similarities and differences between rank order and the regression coefficients from the implicit items (see Table 6.1). On average, participants were aware of which sources they most and least preferred. The implicit and explicit measures agreed for AV crash rates (the most preferred and highest influence) and information from AV companies and friends and family (the least preferred and lowest influence). Participants also agreed on their implicit and explicit assessments of AV near-miss rates, with their assessments appearing in the middle of the pack for both measures. Notably, there was disagreement in influence attributed to messages originating from governing bodies. When asked explicitly, participants rated information originating from federal and state and local governments to be in the lower half (fifth and sixth, respectively), indicating a general disinclination for messages pertaining to AV safety originating from those sources. When measured implicitly, though, official statements from state and local governments and the federal government had the second and third highest influence, respectively. In this survey, it appears that individuals were likely to report that they disregard information coming from the government but do not actually do so when making judgments about AV safety. Conversely, information from advocacy groups and federal vehicle requirements was highly preferred when measured explicitly, but these sources were superseded by government when measured implicitly.

One finding of note is the low influence and preference given for information coming directly from an AV company. Survey respondents did not provide reasoning for their responses, but there was clearly a general disregard for information coming from this source. Likely, there was an issue with credibility; AV companies clearly have a vested interest in AV safety messaging that might be prone to bias. This response from the public is unfortunate because AVs are being developed largely by companies, and, at this point, those companies have the greatest primary knowledge about the capabilities and safety of AVs. It is important to note that this survey measured the influence given to balanced information about AV safety presented as statements from various sources, and it did not measure attitudes toward the source of the data used to justify the statements (aside, perhaps, from the sources directly referencing AV crash and near-miss rates). In the development of broader statements about AV safety, data and research originating from AV companies might still be deemed to be sound by the general public. These results might simply indicate that the public prefers that the data or research be interpreted and reported by other institutions (e.g., government, safety advocacy groups).

Table 6.1. Order of Explicit Rankings and Implicit Influence Attributed to Differing Sources for Messages About AV Safety

Order of Preference	Safety Message Source (measure)	
	Regression Coefficient (implicit)	Rank Order (explicit)
1	AV crash rate	AV crash rate
2	State or local government position	Safety advocacy group position
3	Federal government position	Federal vehicle requirements
4	AV near-miss rate	AV near-miss rate
5	Federal vehicle requirements	Federal government position
6	Safety advocacy group	State or local government position
7	Friends and family members	Friends and family members
8	AV company position	AV company position

NOTE: Order of sources measured implicitly determined by standardized regression coefficients (see Table A.1) from the social judgment analysis. Order of sources measured explicitly determined by mean ranking from the rank-order task.

To summarize the findings from the ALP survey, the public most prefers statements about the safety of AVs from sources that are driven by data and are immediately understandable and relevant (AV crash rates). Their perceptions of safety are then most highly driven by messages from state and local governments, followed by messages from the federal government, regardless of an explicit indication that they have a disinclination for information originating from these sources. Messages from friends and family (who might be uninformed) and AV companies (that might have conflicts of interest) are the least preferred and influence perceptions of AV safety the least of all the sources.

Specific Population Groups

In general, the public is by nature heterogeneous, and specific populations will have different needs and concerns that influence their perceptions of costs and benefits associated with AVs. Some individuals with reduced ability or access to self-operated transportation will emphasize the perceived benefits associated with technology. One particular example is found in older adults—generally a population that reports lower perceived benefits of technology and lower self-efficacy related to technology and that is late to adopt new technologies. In multiple studies, this group has shown an uncharacteristically high positive attitude associated with highly automated driving technologies,¹⁵⁰ including the willingness to pay more for ADAS than

¹⁵⁰ Ericka Rovira, Anne Collins McLaughlin, Richard Pak, and Luke High, “Looking for Age Differences in Self-Driving Vehicles: Examining the Effects of Automation Reliability, Driving Risk, and Physical Impairment on Trust,” *Frontiers in Psychology*, Vol. 10, 2019.

younger adults are.¹⁵¹ Although enthusiasm for ADAS appears high among older adults, it is important to note that older adult attitudes toward fully automated vehicles are more reserved.¹⁵²

Early adopters of technology are another specific group that should be particularly prepared to accept AV technology. The characteristics of these individuals highly promote AV acceptance in light of the constructs and models discussed in this chapter. Early adopters will likely perceive greater benefits and lower risks associated with AVs than will individuals who are less receptive to new technologies. One survey of electric vehicle owners, early adopters of another vehicle technology, showed that they were likely to purchase AVs, though a direct comparison with the general public was not conducted.¹⁵³

Communication Among Stakeholders

Aside from safety statements to the public from certain stakeholders (industry, government officials, or safety researchers and advocates), safety communication occurs between stakeholder groups, primarily downstream from data-generating industry groups to government entities and organizations that conduct safety research and advocacy. The technological expertise of regulators, particularly in relation to individual ADS models, might lag behind that of the developers who spend their days immersed in the nuances of development, testing, and evolution a specific AV technology.

Many of those interviewed—across sectors—expressed such ideas as the following:

[Government personnel] don't have the technical knowledge. Government people have long service and are not corrupted by external pressures, but they are also not current. To understand what's going on with industry you need to be in the industry. Even two years ago, things were different.

In other words,

[t]he expertise does not usually start on the governmental side. AVs are a cutting-edge technology issue. A system that would rely on governmental understanding to advance would not work. The DOT [Department of Transportation] is a reactive regulator of industry, so you can't fault DOT for not being the lead on this issue; that is not how they are structured.

¹⁵¹ Dustin J. Souders, Ryan Best, and Neil Charness, "Valuation of Active Blind Spot Detection Systems by Younger and Older Adults," *Accident Analysis & Prevention*, Vol. 106, 2017.

¹⁵² David W. Eby, Lisa J. Molnar, and Sergiu C. Stanciu, *Older Adults' Attitudes and Opinions About Automated Vehicles: A Literature Review*, Ann Arbor, Mich., and College Station, Tex.: ATLAS Center, ATLAS-2018-26, 2018.

¹⁵³ Rosaria M. Berliner, Scott Hardman, and Gil Tal, "Uncovering Early Adopter's Perceptions and Purchase Intentions of Automated Vehicles: Insights from Early Adopters of Electric Vehicles in California," *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 60, January 2019.

Government officials see themselves playing the roles of communicators and facilitators and of policymakers.¹⁵⁴ As demonstrated by the results from the ALP survey (see Appendix A), government entities have an important role to play in the communication of safety. If government combines technological neutrality and authority to provide distilled safety information for consumption by the general public, industry stakeholders must be able to communicate their measure- and process-based safety approaches to government in ways that are clear and consolidated. According to Bryant Walker Smith:

The public and even regulators lack the resources, technical expertise, capacity, and information to confidently assess the actual performance of systems. Only companies are in a position to do so, albeit under supervision of regulators and others.¹⁵⁵

Government entities and safety advocacy groups have a role as interpreters or intermediaries between the detailed evidence of safety that a company might provide and a general public that wants to know whether AVs meet a threshold, and perhaps (as is the case today with conventional vehicles) how many stars a given model has. With adequate input from industry, the intermediaries can digest and distill the safety information originating with AV companies and then present it to the general public. To illustrate the value of such information and evaluation services, several interviewees pointed to the wide use of Consumer Reports ratings as guides to purchasing for a variety of complex products.

Appraisal

A single broad statement will not provide a blanket solution to the communication of safety to the general public. Different populations have different perspectives and needs. These include varying perceptions of risks and benefits, different information needs, and varying views of the credibility of different sources. Psychological and human factor theory would suggest that statements directed to promote the benefits of AVs are just as important as statements intended to alleviate concerns about the risks and costs associated with AVs. Additionally, addressing one of the variables (e.g., benefits) will influence perceptions of the other (e.g., risks).

Individual preferences might differ, but it seems generally true that simple, distilled statements from trusted sources directly influence public perception of safety and risk. Data from a nationally representative sample are generally clear: Individuals prefer familiar and relevant data-driven arguments to support safety of AVs (i.e., an improvement in the incidence of vehicle crashes) and are influenced more by arguments that come from sources that have some authority and expertise without discernible conflicts of interest (e.g., government rather than friends and

¹⁵⁴ Elaine L. Chao, U.S. Department of Transportation Secretary, remarks as prepared for delivery at the Autonomous Vehicle Symposium, San Francisco, Calif., July 10, 2018.

¹⁵⁵ Bryant Walker Smith, interview with authors, March 19, 2020.

family or AV companies). Improving safety perceptions and mitigating perceptions of risk will promote the acceptance of AV technologies.

As widely mentioned in our interviews with stakeholders, individuals are more accepting of error that originates from humans as opposed to errors from technology. Though the magnitude varies depending on the particulars of any given study on the topic, the general trend is clear: Individuals will not accept AVs until they are safer than human drivers. This expectation relates to the discussion in Chapter 5 regarding thresholds rooted in human performance—notably, how to think about a human driver.

7. Conclusions and Recommendations

Whether AVs are acceptably safe is a matter for assessment and measurement and for communication. Public trust depends on all of these. Improvements are possible in these areas, and our conclusions and recommendations suggest paths forward.

Conclusions

In this report, we explore approaches as tools for stakeholders outside industry to consider as they grapple with asymmetric information about AV safety, and tools for developers to consider as they explore how to demonstrate their evidence of safety and earn trust. Each of the approaches considered in this report has merits and limitations. Those findings are summarized in Table 7.1.

Our research suggests that, as summarized in Table 7.1, the use of process approaches is growing. Comparable, shared, and meaningful quantitative measurements are scarce. This research also indicates that more work—both research and advocacy—is needed to advance development of leading measures (including roadmanship) and meaningful threshold-based approaches (as we discuss in our recommendations).

Meanwhile, relevant processes are proliferating. A greater number of technical standards for AVs is being developed, and best practices are being codified that can form the nucleus of standards development that is more formal and time-consuming. However important they are for structuring and guiding what developers do, processes are not necessarily transparent. That makes communication about them important, as we mention later.

Advancing threshold approaches involves developing a more detailed understanding of human driving. For example, thresholds rooted in human-driving performance are stymied because (1) for a threshold to use leading measures, there are not enough human data for comparison;¹⁵⁶ and (2) for a threshold to use lagging measures, there are not enough AV data for comparison.¹⁵⁷ Through detailed human-driver monitoring in a few areas (today’s “islands of autonomy”¹⁵⁸), it can become possible to assess whether AVs meet a human-driver threshold through roadmanship measures, as discussed in Chapter 5. Thresholds rooted in automated driving systems can be expected to evolve with the technology and the ODD.

¹⁵⁶ This is because there is very little leading-measure material on human drivers, and even less that exists specific to a given AV’s ODD.

¹⁵⁷ This is because it requires considerable time for AVs to accumulate enough miles in a given ODD to generate statistically valid lagging measures—especially of rare events, such as crashes.

¹⁵⁸ Silberg, 2017.

Table 7.1. Approaches to Assessing AV Safety

Approach	Approach Takeaways
Measurement Leading measures Roadmanship Lagging measures	<ul style="list-style-type: none"> • Measurement provides established and easily communicated evidence of safety, but its usefulness can be constrained by (a) a gap between what developers do internally and what can be shared externally for competitive reasons or concerns about how well past data reflect existing safety performance; (b) a lack of meaningful thresholds to contextualize measurement results, resulting in either a lack of comparisons or misleading ones; and (c) immaturity of leading measures, which continue to be developed along with the technology but remain works in progress. • Roadmanship, a leading measure of roadway citizenship, is implicit in leading-measurement development today.
Process Technical standards Government regulation Safety culture	<ul style="list-style-type: none"> • Given the imperfect fit of automated driving systems to regulations developed for conventional vehicles and the constraints on publicly shared, valid measurements, work to adapt existing and develop new processes provides indicators about the quantity and quality of developer attention to safety. Safety cases are a crosscutting process feature.
Thresholds Predicated on human driving Predicated on automated driving system technological performance Predicated on an absolute goal	<ul style="list-style-type: none"> • Thresholds exist in qualitative and quantitative forms. They can be informed by measures or processes or considered on their own. Meeting thresholds is an ongoing, evolving process: There is not a one-time threshold; and thresholds can be internal, for developers, or external, for a broader group of stakeholders. • The lure of comparing automated and human driving is strong. Difficulties arise with implementing human-driving and technology-based thresholds. Absolute thresholds might serve as a transportation-system-wide goal as opposed to an AV-specific one. Thresholds can evolve.

This research also indicates that, given the technical progress of the past couple of years, developers recognize the need to communicate more effectively about the safety implications of what they are doing. Some people consulted for this project pointed to the rise of AV industry coalitions, acknowledging their limitations but suggesting that they provide a basis for communication between industry and government and between industry and the general public. Coalitions that emphasize public communication can affect the thinking of the public both directly and indirectly by influencing government officials.

Communication about AVs and AV safety involves more actors than developers. The research found that the public considers the provenance of different kinds of information and messages about AV safety, having the most confidence in information provided by government and the least in company statements. Regardless of the source, AV communication should factor in the difficulty that people have in gauging risk accurately.

Progress will not come from selecting the best approach—there is no such thing. None of these approaches functions on its own. Evidence of acceptable levels of safety draws on multiple approaches, as illustrated in Figure 7.1. As one interviewee opined, “If there are multiple ways to demonstrate that it is safe enough to pull out a [safety] driver, a suite of them should be used.” A diversity of approaches has been associated with progress in AV development to date. As data and technology improve, different mixes of approaches will become practical.

Figure 7.1. How Approaches Come Together to Show Evidence of and Support Communication About Acceptable Safety of AVs

Process or Measure		Threshold		Statement
AV hard-braking rate	is	below 1 per million VMT	as	communicated by AV developer
AV fatal crash rate per 100 million VMT	is	lower than that of the average human driver	as	communicated by government statistics
AV safety case demonstrates meeting safety standards	showing that	the technology is as safe as possible	as	communicated by a safety advocacy group

As discussed in our prior research, some degree of comparability is important; however, a menu of approaches might provide a path to balancing comparability with differentiation, expanding on the kind of framework discussed in the earlier report. As one interviewee observed:

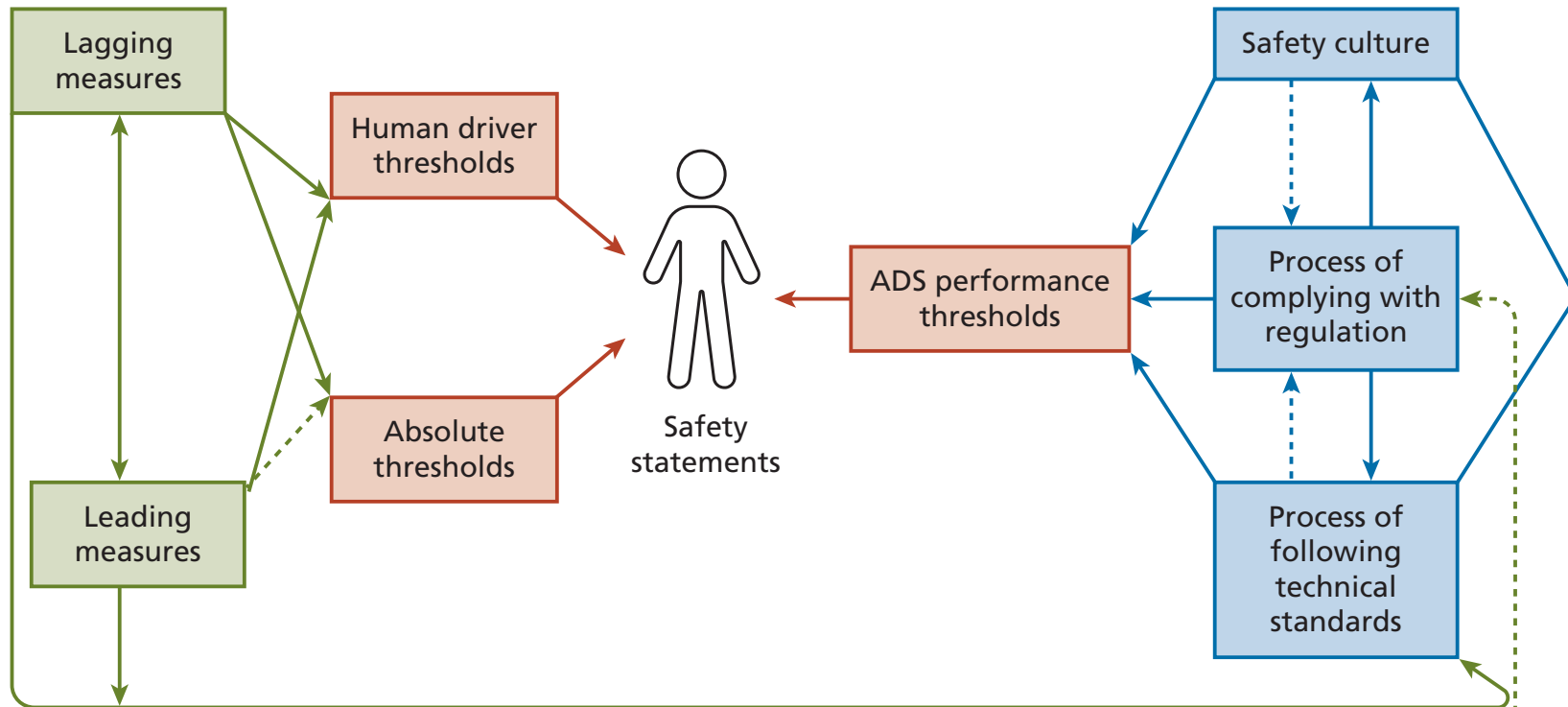
Ask multiple engineers about safety, and they may give you different answers, but they will be related. What are the different dimensions of safety argumentation? First is definitions. What does safety mean for the vehicle in a given environment—the subsystems, components, universal behavior, scenario-specific behaviors, etc.? Definitions could differ by manufacturer, but each wants to define what works for them. In addition to safety definition, what is the technology design—what is the architecture that generates specified safety behavior and avoids unsafe behavior? Third, technical implementation (as distinguished from design). Fourth, the development process—which should be rigorous, as evidenced by artifacts, so that the manufacturer can claim issues were addressed and mitigated, as well as documented. Fifth, wealth of evidence—verification and validation, documentation—show how implementation meets targets. Sixth, operational procedures—maintenance, training, etc. If these things are addressed, companies will create safe vehicles, even if they don’t present information the same way. Any argument for safety would include all six; similar across companies in touching on each dimension but content differing across [original equipment manufacturers].

The framework outlined by this interviewee engages multiple approaches and argues implicitly that doing so is a best practice. It is also, in effect, an individual’s recipe for a safety case (which we discuss later).

How Approaches Build on Each Other

Each of the approaches outlined in Table 7.1 builds on, supports, and contributes to the evidence used in others. Interviewees took pains to discuss the dense web of relationships present, as illustrated in Figure 7.2.

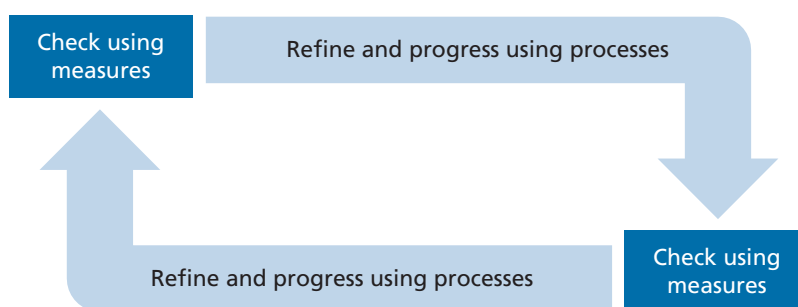
Figure 7.2. How Approaches Build on and Provide Evidence in Support of Each Other



NOTES: The green boxes fall within Safety as a Measurement, the blue boxes within Safety as a Process. The orange boxes represent options within Safety as a Threshold. An arrow indicates the contribution of the originating shape to evidence of safety for the destination shape. An unbroken line indicates a clear relationship. A dashed line indicates that the origin contributes to the destination box only in certain circumstances (e.g., with certain regulation).

Perhaps the strongest and most mutually reinforcing relationship is the one between process approaches and measurement approaches (Figure 7.3). It can be argued that measurement approaches are rooted in a specific time: They show what the vehicle’s event rate per VMT was for a given number miles. Measures show how well the AV has handled the scenarios and the world it has encountered in those miles. This is true for lagging and (especially) for leading measures.¹⁵⁹ Process speaks to the AV’s ability to handle what it has not yet encountered—evidence of whether the vehicle can handle everything for the next given number of miles.¹⁶⁰ And when the vehicle has driven those miles, the measures would again provide a check on the adequacy of the processes used. As this process continues, evidence of safety becomes a staircase to achieving a threshold.

Figure 7.3. How Process and Measurement Approaches Build on and Support Each Other



Interviewees remarked on these relationships, sometimes in response to being asked about how they weighed different kinds of evidence (statistical analysis, processes, expert opinion). The following examples cited by interviewees capture the limitations of available evidence and reinforce what each type of approach can offer the other.

A lot of standards around process leave a lot open to interpretation. They depend on expert opinion or other subjective measure of whether an entity is complying. We need objective measures about outcomes. The end check will be statistical evidence. Expert opinion helps move process—what statistics should we look at, are we implementing that part of standards—but it shouldn’t be part of the ultimate measurement of whether AVs are safe or not safe.

The reality is that, without metrics, where is the oversight to make sure people meet those metrics? Cars are self-certifying to what? A system that’s not in place. NHTSA uses recalls but needs metrics to actually evaluate that they are safe. . . . People are designing with no metric in place. We see that with forward collision systems. What is an effective system, how do you evaluate and guide

¹⁵⁹ This is because it is unlikely that the AV will have traveled enough miles to generate statistically valid lagging measures.

¹⁶⁰ For example, one could consider whether the validation, verification, and due diligence were sufficient to ensure desirable behavior for mileage and scenarios not yet encountered in the real world.

manufacturers? With forward collision warning, NHTSA has developed testing protocols. With level 2, [there is] no basic guidance, let alone testing protocols, but we don't have guidance for driver monitoring and other aspects.

Such comments recognize that measures are a retrospective accounting and process is a prospective expectation. Ironically, virtually everyone interviewed for this project—who were all experts of some kind—dismissed the value of expert opinion as a compelling form of evidence to determine whether AVs are safe enough when compared with statistics and demonstrated compliance with standards.¹⁶¹

As Figure 7.1 shows, measures can support thresholds and processes, notably those predicated on human driving and on absolute goals, which can then be communicated to the general public by AV companies, government, and safety advocates.

Processes are generally mutually reinforcing and are used to show evidence of meeting a threshold predicated on the ADS's performance. Although aspects of processes (such as use or compliance) are likely to be communicated by a company, processes (other than compliance with regulation) are less likely to be discussed with the general public by government and safety advocates. This is because process information likely comes unilaterally from an AV company—unlike measures, which might be reported by police and other entities. So although government and advocates might discuss processes with companies, these two groups are unlikely to make statements about process to the general public beyond making public what an AV company has said (e.g., the public Voluntary Safety Self-Assessments and applications for exemption).¹⁶²

Safety Cases

As noted in Chapter 4, safety cases have been around for a long time. Their application to software systems has been growing since the 1990s.¹⁶³ Several factors have combined to elevate their importance for AV development. These go from a growing role in the context of proliferating technical standards (as exemplified by their centrality to the new standard, UL4600) to the recognition that such broad approaches as the Federal Aviation Administration's Safety Management System, which centers on safety cases, have value in AVs. Although safety cases began as a way of organizing and documenting engineering and other development activity, they can have broader scope. Their use in UL4600 illustrates how they can extend to other activities that convey safety culture; accordingly, safety cases should have appeal as bases for a broader

¹⁶¹ *Expert opinion* would be defined as evidence based on a trusted individual or institution's judgment or endorsement. Experts can be federal, state, and local regulatory bodies; safety advocacy groups; industry statements; academics; local dealerships; or even friends or family members.

¹⁶² Safety culture might be an exception here. Companies discuss their safety culture with the general public (e.g., Aurora's publications and discussions on its safety culture practices), and in the past, the National Transportation Safety Board and safety advocates have cited safety culture as a contributor to or protector against crashes. Chris Urmson, "Putting Safety into Practice: Aurora's Safety Approach," Aurora blog post, October 2, 2019; Steve Finlay, "Uber Improves 'Safety Culture' in Aftermath of Fatal AV Accident," WardsAuto, August 7, 2019.

¹⁶³ Bloomfield and Bishop, 2010.

menu of communication to a broad array of stakeholders. What specifically is communicated however, is an open question.¹⁶⁴ AV developers can help offset the information asymmetry previously noted, but their proprietary concerns are expected to continue to limit what they make public.

Leading Measures and Roadmanship

During our interviews, we observed that the terms “lagging” and “leading” measures are in wider use among key stakeholders than they were during our prior research. The persuasiveness of leading and lagging measures, noted in Chapter 3, is illustrated by the survey conducted for this project (discussed in Chapter 6 and Appendix A), which showed that key leading measures (such as the near-miss rate) and lagging measures (such as the average crash rate) influence the general public’s perception of AV safety. Lagging measures were the source of information with the greatest influence in predicting public perceptions of AV safety. That likely reflects decades of previous use, the clarity of measure, and the personal relevance to real-world outcomes for the survey respondents. These correspond to human-driving experiences and illustrate the challenge (also discussed in Chapter 6) of communicating more-abstract measurements.

Roadmanship, the term and concept introduced in *Measuring Automated Vehicle Safety: Forging a Framework*,¹⁶⁵ was recognized by several in the industry, even without there having been systematic advocacy for it. Interviews showed that discussions of leading measures tend to center on roadmanship even if that term was not used—the more general term, “leading measures,” effectively refers to roadmanship measures (due partially to a dearth of other options). Measurement uncertainties and the many pointers toward the need for an integrated approach suggest that roadmanship has a role—a role that could be more obvious with a champion, perhaps a consortium, such as AVSC.

Mapping Agreement and Disagreement

There is no prospect of agreement among stakeholders without common understanding that there is some baseline level of risk.¹⁶⁶ But even with that understanding, views differ on how much risk is acceptable. One interviewee explained that his organization “will never be satisfied with safety, because as long as there are crashes or fatalities, things could be safer,” expressing concern that the “message directed towards the user is that accidents are unavoidable.” Although people aligned with safety advocacy might be more insistent on that perspective, cautioning against complacency, the

¹⁶⁴ Safety cases in other contexts historically have not been published, in part because of a range of considerations from security to proprietary content. See Bloomfield and Bishop, 2010.

¹⁶⁵ Fraade-Blanar et al., 2018.

¹⁶⁶ Alain Kornhauser highlighted the problem succinctly: “Since safety is really a relative perception, rather than an absolute fact, fighting relative perception with facts is usually futile and often counterproductive.” Alain Kornhauser, “8.25-Birthday-0612420,” *Smart Driving Cars*, newsletter, June 12, 2020.

expectation of continual improvement for automated driving systems is broadly held across stakeholder groups as a practice that is associated with software specifically and technology generally. Hence the argument is about what the floor should be—in effect, a question of threshold.

There is broad agreement across stakeholders on the challenge of communicating about risk. As discussed in Chapter 5, people interviewed were invariably drawn to some kind of comparison between automated driving systems and human drivers, even if they were articulate about the limitations of such comparisons and even though they differed in the kinds of comparisons they declared most suitable. The lure (some might say, the logic) of comparing automated driving systems with human drivers, however fraught, is thus an area of agreement. The analysis presented in Chapter 5 shows that it matters how such a comparison is made; it is possible to do better than a comparison to the national average human driver, but corresponding data are needed. Agreement is also needed on what it means to be on par with the higher threshold of a safe human driver. Developer agreement on a threshold of a better-than-average, or safe, driver would foster comparability and appeal to other kinds of stakeholders.

Stakeholder Ecosystem

AV safety is a political and a substantive issue. That is why communication (discussed in the next section) is central to the question of whether AVs are acceptably safe. Developers and the AV industry depend on either facilitation or removal of barriers by government at different levels, and on public trust, without which there is no market. In turn, what both developers and government do is monitored by research and advocacy groups on one hand and the general public on the other. The absence of shared information and common approaches to assessing and communicating about safety across these groups generates friction. It also hinders agreement on what it means for AVs to be acceptably safe.

AVs Are Developed by Businesses

The fact that AVs are developed by businesses (both start-ups dedicated to automation and more-established companies with histories in either information technology or motor vehicles) is key context for understanding disparate stakeholder views.¹⁶⁷ There is broad understanding that, as one interviewee said, developers are in a “race to get safe enough . . . or go out of business.” Of course, society reaps the benefits of innovation because people have seen a business opportunity and have been willing to invest and to incur the risk of business failure. As one person outside the industry observed:

¹⁶⁷ Past safety-related incidents involving motor vehicles and other kinds of products have motivated both the growth of entities (within government and as nonprofits in the private sector) focused on consumer protection and their baseline skepticism (and, sometimes, outright adversarial posture toward industry). Such attitudes draw from investigation findings that companies demonstrated poor internal coordination and inadequate attention to safety.

I do hope and expect that AVs will be dramatically safer than even the best human drivers and eliminate fatalities on the road. If the great economic engine and profit opportunity where humans are replaced by robot drivers [are] not saving a lot of lives, it would be a failed investment relative to the opportunity.

An individual within the AV industry stated:

[What is acceptably safe is] a business question. The whole AV business recognizes how important it is to deliver a safe product. Look at aviation—safety is essential. Employees own the safety. There is an innate push for safety, which might lead to more-stringent levels than what is required based on legal or societal grounds. But if you require AVs to be too safe then you forgo benefits to society by forgoing better technology.

This project points to ways in which AV developers can demonstrate and communicate better about AV safety.¹⁶⁸

A challenge of assessing whether and when AVs are acceptably safe is that, with complex products, such as AVs, the producers inevitably know more than the public or others can understand, a situation known as *information asymmetry*. Although that reality has been a motivation for regulation generally, regulators inevitably lag behind industry when it comes to understanding how the relevant science or technology works, another reality that was mentioned by a wide variety of stakeholders in interviews. Similar circumstances have been seen with conventional motor vehicles, pharmaceuticals, medical devices, and other complex products that can affect health and safety.¹⁶⁹

With AVs, the dynamic nature of the technologies and asymmetric understanding have led to a situation in which most of the stakeholders interviewed described regulations as providing only a floor for the protection of safety. Interviews also documented the continuing interest outside the industry in regulation of some kind as a response to asymmetric information and safety concerns.

Information asymmetry combined with the inability to demonstrate safety conclusively by testing and measurement leads all kinds of stakeholders to spotlight safety culture, which animates the process category. Reputation and good will contribute to trust, and, for a complex

¹⁶⁸ Note that our 2018 report called for demonstrations that convey ADS progress and proficiency. See Fraade-Blanar et al., 2018.

¹⁶⁹ These dynamics have fueled the study of regulatory economics and policy (see Viscusi, 2006). For example, the FDA regulates both products and manufacturing processes for pharmaceuticals and medical devices, and it occasionally requires recalls. FDA, “Medical Devices,” webpage, undated; FDA, “Digital Health,” webpage, August 27, 2020.

Over the past few decades, that process has increasingly addressed the introduction and use of software in such systems, including advancing a recent Digital Health Innovation Action Plan and within it a Software Precertification Pilot Program. Although the “is proposing that software products from precertified companies would continue to meet the same safety and effectiveness standard that the agency expects for products that have followed the traditional path to market” (FDA, “Digital Health Software Precertification (Pre-Cert) Program,” webpage, July 18, 2019a), it is putting a spotlight on processes (as discussed in Chapter 4): “This proposed approach aims to look first at the software developer or digital health technology developer, rather than primarily at the product, which is what we currently do for traditional medical devices” (Agata Dabrowska and Susan Thaul, *How FDA Approves Drugs and Regulates Their Safety and Effectiveness*, Washington, D.C.: Congressional Research Service, R41983, May 8, 2018).

and effectively opaque kind of product, demonstrating safety culture is broadly seen as essential for developer reputation.¹⁷⁰ As discussed earlier in this chapter and in Chapter 4, safety cases can support both safety culture and the positive trust balance noted earlier.

Our 2018 report called for sharing information associated with adverse incidents as case studies.¹⁷¹ In the absence of such sharing, National Transportation Safety Board investigations appear to play that role. Stakeholders continue to differ, as documented in the team’s interviews, about what kinds of data might be shared, either among developers or between developers and government (let alone the safety research and advocacy community); stakeholders pointed to disengagement reporting in California as providing an illustration of how dependent any data are on circumstances, such as ODD or a specific route. Because of such dependency, interviews with developers even noted potential competitive concerns associated with testing that is specific to demonstrating safety. As one put it, “How do companies meet competing goals of collaborating around safety and preserving proprietary advantage that isn’t lost in the testing process?” But as another lamented, it is “hard to establish safety when it is all customized data.” A third expressed the kind of impatience common outside the industry: “There is enough experience in the last five years that there should be some common measures.”

There might be a self-correcting aspect as the industry consolidates, resulting in fewer developers being concerned about protecting proprietary interests—but not enough to obviate the concern during this important phase of technology and industry maturation. Two people outside the industry summed up the situation as follows:

Looking holistically, how we get at what measures there are [in common use] will take a lot of work by people that don’t come to the table with the same understanding of the technology or understanding of the different use cases. Developers and [original equipment manufacturers] have a certain perspective; regulators come from a different perspective. It would take a lot of discussion to get to where all can agree on a particular set of guidelines. And that’s just two kinds of interested parties; we need others—advocates, users, others.

Operators [i.e., safety drivers] and safety culture are useful to show, but at the end of the day, [government] just doesn’t know. Today technology is good, but no one has clear direction, no one has a clear swim lane. The federal government cannot say, “This [policy arena] is ours,” but then be hands off. It cannot make [AV safety assessment] a pure self-certification approach. There are so many groups getting involved; we need some sort of protocol. . . . Who is the gatekeeper?

Against the backdrop of such continuing disagreement across categories of stakeholders, NHTSA launched the new AV TEST Initiative in June 2020,¹⁷² promising a public-facing

¹⁷⁰ Or, as one interviewee put it, “[T]he only thing the AV sector can do to gain good will is to reveal the people behind the industry, get people talking.”

¹⁷¹ Fraade-Blanar et al., 2018.

¹⁷² NHTSA, “New Test Tracking Tool,” webpage, undated-a.

platform for data contributed voluntarily by developers and by states and localities that host testing. Stakeholders interviewed for the project struggled with the idea of a third-party review of key information—the merits of independence were recognized, but the feasibility of achieving it poses numerous challenges, beginning with how information asymmetry limits the understanding of outside evaluators. It will be interesting to see how developers respond to the call for independent assessment in the new UL4600.

Liability

Many stakeholders interviewed for this project remarked on how regulation is complemented by liability (and insurance, especially for those operating AV fleets) in shaping the incentives for how AV businesses make decisions that bear on safety.¹⁷³ The ex post liability of developers for their products (or the liability of service providers operating AVs if they knowingly choose vehicles known not to be the safest) is broadly acknowledged as influencing the evolution of the AV industry and of AVs. As one interviewee summarized the situation, “[E]xisting liability frameworks can help encourage safe behavior by these companies.” Another added emphasis:

Perhaps these companies have concluded, as have others, that existing tort law, while messy, is nonetheless familiar and ultimately tolerable. Perhaps they are wary of opening cans of worms that could wriggle far beyond their control. Perhaps they are waiting until they know what they want and have the political power to achieve it. Perhaps they do not want to make needless opponents of automated driving. Perhaps they understand the dissonance between, on one hand, promising safety and asking for trust and, on the other hand, promising destruction and asking for immunity. And perhaps they, too, see automated driving as a technological solution, for ultimately the best way to reduce liability is to reduce injury.

Excessive liability costs can lead companies to exit a market,¹⁷⁴ which might result in remaining companies not being the safest but rather the ones with the deepest pockets. The prospect of high costs of many kinds is already reflected in industry consolidation. The ability of AV developers to assess and communicate convincingly about the safety of their automated driving systems will help to set the stage for the (further) development of these and other mechanisms. If what is regulated and communicated continues to seem too meager to some stakeholders, the project’s interviews suggest that those stakeholders will look to liability as a more visible influence.

Improving Communication About AV Safety

All of the stakeholders interviewed do and will communicate about AV safety. All acknowledged that communicating about AV safety to the general public is difficult and could be

¹⁷³ Zev Winkelman, Maya Buenaventura, James M. Anderson, Nahom M. Beyene, Pavan Katkar, and Gregory Cyril Baumann, *When Autonomous Vehicles Are Hacked, Who Is Liable?* Santa Monica, Calif.: RAND Corporation, RR-2654-RC, 2019.

¹⁷⁴ Viscusi, 2006.

improved. What is communicated—and how—will shape opinions about whether AVs are acceptably safe. Two interview quotations speak to the need:

People will need to be taught about what to expect of encounters with automation. The technical underpinnings [of AVs] complicate [and lead to] misunderstanding. There is a mismatch between where technology really is and people's expectations of the technology.

How do you take something complicated and simplify it for consumer audiences? Consumer Reports has information about the best cars, which have the best gas mileage. Consumers will be looking for that. But 80 percent will just be looking for, "is it safe enough?" It's not happening now [with AVs], but most people will assume that things are safe enough (like they do a water park). Some people will really want to understand it, maybe early adopters. They may be critical of it; they may be excited about it. You see this in Tesla owners, who are excited about electric vehicles.

What is measured and communicated might differ before and after deployment.¹⁷⁵

Acknowledging the value of simple messages, one interviewee forecast a kind of "restaurant rating" system with three simple tiers (safe, safer, safest) that summarize what can be known without getting into the details.

Those ratings will be masking a moving, numerical metric that is evolving over time as AVs get safer. You don't want to say that last year's rating was not safe but next year's is.

Such simplicity echoes the star-rating systems that already exist.¹⁷⁶ As one interviewee explained,

For the general public, the crash ratings from NHTSA—they're just stars. That's what the public wants: government institutions providing simple information.

This kind of high-level rating could float atop changes in the underlying performance levels, making it dynamic rather than static.

Many people we interviewed found room for improvement in communication about AVs and how safe they are. Interviewees both in and outside the industry expressed concerns about the perils of publicizing only problems. People who study risk perception understand that

¹⁷⁵ For example, testing with a safety driver could be associated with communication about backup human oversight.

¹⁷⁶ These are provided by NHTSA and IIHS. According to NHTSA, "The National Highway Traffic Safety Administration's New Car Assessment Program (NCAP) created the 5-Star Safety Ratings Program to provide consumers with information about the crash protection and rollover safety of new vehicles beyond what is required by Federal law. One star is the lowest rating; five stars is the highest. More stars equal safer cars" (NHTSA, "Ratings," webpage, undated-b). NHTSA announced plans to revive NCAP in late 2019 and requested comment on its specific information-gathering proposal in mid-2020. NHTSA, "NHTSA Announces Coming Upgrades to New Car Assessment Program," news release, October 16, 2019b; NHTSA, "Agency Information Collection Activities; Notice and Request for Comment; Government 5-Star Safety Ratings Label Consumer Research," *Federal Register*, Vol. 85, No. 23598, April 28, 2020b.

“[s]ubstantial publicity regarding a risk may lead to risk overestimation,” because “[w]hat is being publicized is the numerator of the risk calculation rather than the denominator and the denominator or the overall risk frequency.”¹⁷⁷ This is consistent with the discussion of Safety I and Safety II in Chapter 3, and consistent with the discussion of common biases in Chapter 6.

There are opportunities for all stakeholders to do better at conveying potential AV benefits. One interviewee took the long view in referring to how AVs might fit into a future set of transportation options, sketching out a narrative for the public:

We want to do what we can to achieve public safety and accessibility to transportation. At some point, vehicle ownership will decrease. We will have the ability to live in a world where [technology] will plot our daily itinerary. We might take an AV to the train and then get a bike or pod to the office. The future is more about mobility on demand and access to it . . . more mobility as a service.

Setting aside the details of such a narrative, there is benefit in considering AVs and their safety in a broader mobility context. More narrowly, where testing is happening, different kinds of stakeholders have opportunities to explain to the public what they might see or experience, and what they should know and understand about AVs.

Given limitations to how people process information about risk, as discussed in Chapter 6, several interviewees referred to the benefit of exposure: Do people see AVs on the road? Have people taken a ride in an AV? The slow expansion of testing in a large country, such as the United States, that presents myriad ODDs suggests that it will be long time before most people see AVs in their neighborhoods.¹⁷⁸ Some kind of literal road show—an effort to bring AVs into communities for brief but publicized intervals to at least be observed in action—could help more people think about AVs more concretely.

Next Steps

The COVID-19 pandemic illustrates the tensions and choices between being reactive and proactive in the face of risk, and the challenges of both understanding and communicating about a dynamic phenomenon affecting public health. It has focused global attention on the interplay of public and private sectors in managing risk for the public good. For AVs, one interviewee suggested an ideal for development might be antifragility.¹⁷⁹ An antifragile automated system would be able to respond better to new challenges after encountering a difficult situation than it was able to before.¹⁸⁰ Associated assessment challenges might dwarf those of today, but meeting

¹⁷⁷ Viscusi, 2006.

¹⁷⁸ A survey by Partners for Automated Vehicle Education also connected awareness with attitudes (Partners for Automated Vehicle Education, “PAVE Poll: Americans Wary of AVs but Say Education and Experience with Technology Can Build Trust,” webpage, undated; Partners for Automated Vehicle Education, “PAVE Poll: Fact Sheet,” May 2020).

¹⁷⁹ Nassim Nicholas Taleb, *Antifragile: Things That Gain from Disorder*, New York: Random House, 2014.

¹⁸⁰ By extension, developers need to seek problem situations in testing AVs.

those of today is a necessary first step. In this final section, we offer recommendations, drawn from the analysis presented in this report, for different groups of stakeholders.

Recommendations for AV Developers

Developers should explore how to leverage a mix of approaches as part of their safety assessment and communication. Doing so will support a common framework for measurement. In addition, the approaches that are presented and analyzed in this report can contribute to better assessment and communication about AV progress toward being acceptably safe. Furthermore, these approaches can provide common scaffolding for assessing and communicating about AV safety. Our survey demonstrates the importance of how communications are structured, particularly with the general public, and it suggests that communications about AV safety should be data-driven but not too complex, speak to real-world scenarios, and come from authoritative bodies that have no obvious bias (i.e., government at different levels). Although the opportunity is clear for government entities, the data will come from the developers.

AV developers should collaborate on developing publicly accessible versions of their safety cases that are comprehensible to laypeople, perhaps through one of the coalitions that has emerged. The Voluntary Safety Self-Assessments are idiosyncratic, uneven in treatment of safety, and promotional for those who complete them, but the creation of a standard template for a publicly shared safety case would foster transparency and comparability—it could be endorsed by an industry group as a best practice. Given differences across ODDs, use cases, business models, technology, and so on, developers should forge uniform and transparent approaches to presenting evidence for meeting safety thresholds (or supporting other approaches). One aspect could be agreement on using a safe or better-than-average driver as a threshold.

AV developers and the larger AV research community should continue to advance and integrate leading measures, including roadmanship. Again, a consortium could help, including by becoming a champion for advancing roadmanship measures.

AV developers should collaborate with state and local leaders to bring their vehicles into communities around the country. This collaboration could include demonstrations of how AVs operate, things they can do that might improve on human perception and reaction capabilities, and the nature and implications of ODDs.

Recommendations for the Federal Government

The U.S. Department of Transportation should support further research into and data about human drivers to enable the kinds of comparisons broadly sought, working with other federal agencies, industry, and/or state departments of transportation. Examples include understanding how people drive in different ODDs, near misses in different ODDs, and similar assessments to enable (1) better comparisons of automated driving systems with human drivers and (2) understanding of how to characterize a safe, rather than an average, human driver. Given the broad diffusion of artificial-intelligence-based “autonomy” occurring in the economy, there is a potential for synergy or at least cross-domain learning between automated driving systems and

other kinds of automated systems, which is the kind of possibility that goes well beyond the interests of a company but would appeal to the research community.

The federal government should support more research into safety assessment options.

As long as the gold standard among approaches is measurement, and as long as available leading measurements remain deficient (and lagging measures unattainable), there is need for research into advancing measurement options and abilities and into processes overall. Some of the activity associated with standards-setting might engage applied research, but a benefit of government-sponsored research is the expectation of published results. As the industry matures, and as more data sets are shared by more developers (e.g., Waymo and Lyft¹⁸¹), the prospects for research that is useful in establishing measurements should improve.

¹⁸¹ Waymo, “Waymo Open Dataset,” data set, undated; Lyft, “Level 5 Open Data: Advancing Self-Driving Technology, Together,” data set, undated.

Appendix A. American Life Panel Survey

When making a decision through the integration of information from different sources, the identity of the sources themselves can independently affect individuals' decisionmaking. To estimate the weights attributed to differing sources of safety messages pertaining to the safety of AVs, we conducted an Internet-based study using the RAND American Life Panel (ALP),¹⁸² a nationally representative Internet sample of more than 6,000 individuals. ALP surveys are administered to panel members digitally. Those without Internet-connected devices are provided with equipment. For additional information on the ALP, please consult the technical description.¹⁸³

Methods

In April 2020, a survey instrument created by the authors of this report was administered online to ALP participants. The raw data will be made available online after an embargo period of one year.

Participants

Data were collected from 2,203 participants (ages 23–90 years, mean age = 58.42, standard deviation age = 13.87; 56.7 percent female). Panel members are recruited primarily through probability sampling with a separate and smaller secondary convenience sample mainly used for pilot testing. Members from the convenience sample were used for pilot testing ($n = 50$) of the study while the probability sample was used for the study population ($n = 2,203$) used in the subsequent analyses. As with most scientific survey samples, the composition of the ALP is not a precise match with the reference population (i.e., U.S. adults). Therefore, weights were used to correct for sampling error. Using demographic data from the Current Population Survey as a benchmark,¹⁸⁴ ranked weights were calculated to match population proportions along five two-way distributions: gender by age, gender by ethnicity, gender by education, gender by household income, and household income by number of household members. These weights were included into the regression analyses to correct for demographic differences between the sample and the U.S. population (e.g., ALP participants in our sample are older and more White than the U.S. population).

¹⁸² RAND Corporation, American Life Panel, “Welcome to the ALP Data Pages,” webpage, undated.

¹⁸³ Michael Pollard and Matthew D. Baird, *The RAND American Life Panel: Technical Description*, Santa Monica, Calif.: RAND Corporation, RR-1651, 2017.

¹⁸⁴ U.S. Census Bureau and U.S. Bureau of Labor Statistics, “Current Population Survey (CPS),” webpage, undated.

Materials

A survey instrument was created employing 20 survey items to implicitly estimate the relative influence of safety information originating from eight different sources: crash rates of AVs compared with those of human drivers, average near-miss rates for AVs compared with those of human drivers, federal vehicle requirements, federal government official position, state or local government official position, AV company official position, safety advocacy group official position, and friend or family recommendation. For our survey, we considered these eight sources to be similar in their ability to provide safety information to the public, even though one can identify two separate groups: quantitative sources, in which something is measured or a threshold is met (crash rates, near-miss rates, and federal vehicle requirements), and qualitative sources, in which an official statement or opinion is provided (the other five sources). We chose to combine these sources, even though there are scenarios in which one or more of the quantitative message sources' information might be communicated to the public through one or more of the qualitative sources. We kept these as separate groups because these sources of information might not always agree with one another, and we wanted to allow for that information pattern in the structure of the survey (as we will describe). For example, in states where AV testing is being conducted, the AV companies and local governments have communicated to the public that AVs are safe even though reliable data on rates of crashes and near misses are unable to provide information about safety.








The survey employed a social judgment analysis design to determine the influence of positive, negative, or absent information from various sources on the perceived safety of AVs.¹⁸⁵ *Social judgment analysis* is a research technique that uses multiple regressions to analyze survey data to identify the latent judgment preferences (in our study, the influence of information on perception of safety) held by individuals.¹⁸⁶ For each of the 20 implicit items used in the social judgment analysis, participants were shown a table listing the eight safety message sources and what information they were providing. Any source could provide one of five messages within any given item (see Figure A.1): a graphic of two green “thumbs up” indicating that the source strongly shows that AVs are safe, a graphic of one green “thumb up” indicating that the source mostly shows that AVs are safe, text stating “no information” indicating that the source has not provided information, a graphic of one red “thumb down” indicating that the source mostly shows that AVs are unsafe, and a graphic of two red “thumbs down” indicating that the source strongly shows that AVs are unsafe. After considering the information provided by the eight sources, participants gave ratings of their perception of the safety of AVs using a slider on a

¹⁸⁵ Joette Stefl-Mabry, “A Social Judgment Analysis of Information Source Preference Profiles: An Exploratory Study to Empirically Represent Media Selection Patterns,” *Journal of the American Society for Information Science and Technology*, Vol. 54, No. 9, 2003.

¹⁸⁶ Ray W. Cooksey, “The Methodology of Social Judgement Theory,” *Thinking & Reasoning*, Vol. 2, No. 2–3, 1996.

scale ranging from “not at all safe” to “completely safe.” The scale also included labeled markers at positions corresponding to 25 percent, 50 percent, and 75 percent of the scale length, labeled “a little safe,” “generally safe,” and “very safe,” respectively.

Figure A.1. Sample Survey Prompt

Source of evidence	Evidence shows that AVs are safe
Average AV crash rate	
Average near-miss crash rate	
Federal vehicle requirements	
Federal government official position	
State or local government official position	
AV company's official position	No information
Safety advocacy group's official position	
Friends or family members	

NOTE: Two green thumbs up indicates that the source strongly shows that AVs are safe; one green thumb up indicates that the source mostly shows that AVs are safe; one red thumb down indicates that the source mostly shows that AVs are unsafe; and two red thumbs down indicates that the source strongly shows that AVs are unsafe.

In addition to the implicit items used to estimate the influence of safety messages from the different sources, participants were asked to explicitly rank order their preference for the message sources. Participants also answered four additional items pertaining to their comfort with technology and AVs and two items on their experience with ride-share transportation.

Procedure

After consenting to participate in the study, participants first read a screen briefly describing AVs and were introduced to the design and purpose of the study: determining how the participant thinks about the safety of AVs after considering different information about AV safety from various sources. Participants were then introduced to the eight different safety message sources, the five different messages about AV safety that could be provided by the sources, and the safety rating scale. After being shown a sample item, participants completed the 20 implicit items, which appeared in a randomized order. Next, participants were asked to rank order their preferences of message sources. Lastly, participants completed the six Likert-scale

survey items. The demographic information used in our analyses is collected quarterly and was on record with the ALP.

Results

For the primary analysis of the implicit items, the data were analyzed in a linear mixed model predicting perceived safety ratings, including (1) participant as a random factor (to account for the repeated measurements within each participant) and (2) the message from each safety message source, age, and gender as fixed factors. The message from each source was coded as –2 for two red thumbs down through 2 for two green thumbs up. As seen in Table A.1, all predictors reached statistical significance. The sample size we used gave sufficient power to discover a small effect. Standardized regression coefficients (β) were calculated by rerunning the model replacing all variables with their z -scores.

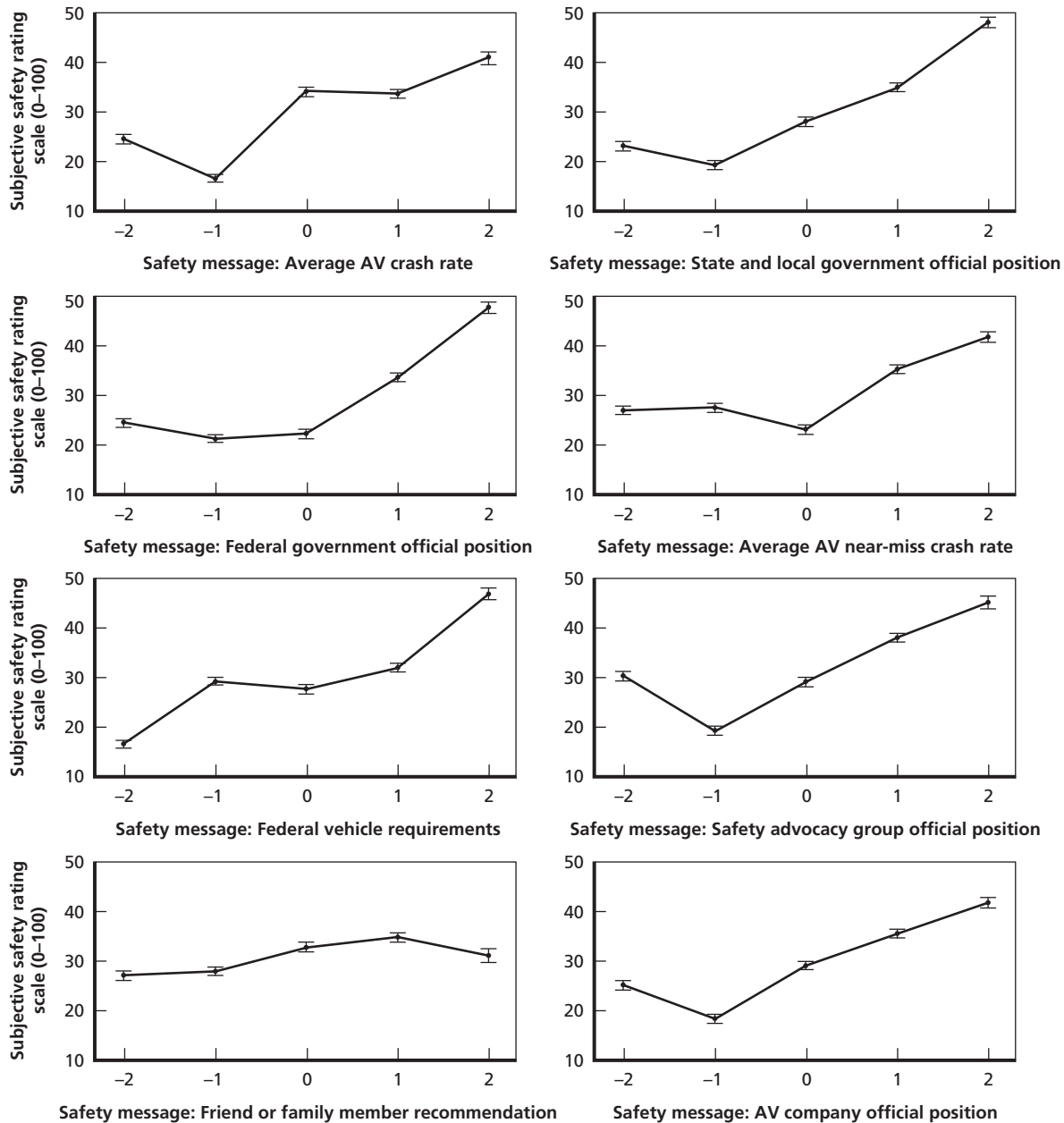
Table A.1. Linear Mixed Model Predicting Safety Ratings Using Safety Message Source, Age, and Gender

Factor	β	b	Standard Error	t	p
Intercept	<0.01	–0.23	10.27	–18.08	<0.001
AV crash rate	0.25	4.46	0.09	52.39	<0.001
State or local government official position	0.22	3.99	0.11	37.92	<0.001
Federal government official position	0.16	2.99	0.08	36.26	<0.001
AV near-miss rate	0.16	2.98	0.09	32.94	<0.001
Federal vehicle requirements	0.10	1.84	0.08	23.73	<0.001
Safety advocacy group official position	0.06	1.08	0.10	10.69	<0.001
Friends and family members	0.05	1.02	0.08	12.75	<0.001
AV company official position	0.03	0.55	0.07	7.30	<0.001
Age	–0.08	–0.14	0.02	–6.76	<0.001
Male	0.09	4.58	0.58	7.97	<0.001

NOTES: The standardized β values reported in this table were calculated as the linear mixed-model coefficients after z -scoring all items. β = standardized coefficient; b = unstandardized coefficient; t = t -value; p = p -value.

As shown in Table A.1, although safety messages from all sources were positively related to perceived safety (i.e., as the message improved from strongly showing AVs were unsafe to strongly showing AVs were safe, the perception of safety of AVs improved; see Figure A.2), the message sources differed in the size of the effect and the weight of their message on the perception of safety. Figure A.2 displays the average perceived safety score (y -axis) for each safety message (x -axis) for each of the eight message sources. Unlike the regression analysis reported in Table A.1, the perceived safety score values presented in each of the eight graphs in Figure A.2 do not control for other variables, including the safety messages from the other seven message sources.

Figure A.2. Perceived Safety Ratings, by Safety Message Source and Content



The message sources “AV crash rate,” “state or local government official position,” “federal government official position,” and “AV near-miss rate” all greatly influenced the perceived safety ratings. Although controlling for the other variables (i.e., age, gender, and the other messages; see Table A.1), each additional step increase in the message from strongly negative to strongly positive increased the perceived safety score by between approximately 3 percent and 4.5 percent. “Federal vehicle requirements” were also a moderately strong predictor, increasing the perceived safety score by 1.84 percent with each step increase in message positivity. “Safety

advocacy group official position” and “friends and family” each increased the perceived safety score by approximately 1 percent with each step increase, although “AV company official position” had the lowest effect with a 0.55-percent increase with each step increase in message positivity. Controlling for the various messages from the safety message sources, increased age was associated with lower perceived safety of AVs, and males rated the perceived safety of AVs higher than females did.

Interestingly, the rank ordering of preference for message sources differed compared with the ordering provided by the standardized regression coefficients. When asked to rank order their preferences, individuals on average provided the following order: AV crash rate, safety advocacy group official position, federal vehicle requirements, AV near-miss rate, federal government official position, state or local government official position, friends and family members, and AV company official position. In general, participants’ preferences matched well to the highest and lowest regression coefficients estimated from the implicit items. But assuming that individuals are more influenced by their preferred sources of safety messages, participants in our sample overestimated the influence of information from safety advocacy groups and federal vehicle requirements while underestimating the influence of information from all governmental official positions.

Individual Cases and Comparisons

Individual implicit items were constructed to reflect particular information patterns that might reflect real-world messages about AVs. Overall, the perception of safety across all implicit items was low. Across the entire data set, the mean safety rating was 31.5 (standard deviation = 13.29) out of a 100-point scale. Separated by item (see Table A.1), the only item that crossed the midpoint on average was question 19, which listed all safety message sources as showing strong support for AV safety.

Agreement Among All Sources, Both Positive and Negative

To ensure that increasingly positive safety messages resulted in higher perceptions of safety (i.e., a manipulation check) and in an effort to gauge the overall range of perceived safety for the survey design, two items were added for which there was consensus among all sources that AVs were unsafe (item 17) or that AVs were safe (item 19; see Table A.2). Item 17 was the lowest-rated item, significantly lower than item 16 (all sources provided messages that AVs were unsafe except for friends and family, who provided a message that AVs were safe), which was the item with the next lowest mean rating ($F[1, 2,153] = 359.00, p < 0.001$). Item 19 was the highest-rated item by a large margin, 28.3 points greater than the next highest item (which was item 14: strong messages from government that AVs were safe, supporting messages from AV companies and advocacy groups that AVs are safe, contrasted with information from average near-misses that AVs are not safe). This difference was highly significant ($F[1, 2,164] = 3,052.16, p < 0.001$).

Table A.2. Means and Standard Deviations of Perceived Safety, by Survey Item

Situation Number	Safety Message Sources								Ratings	
	CR	NM	FVR	FG	SLG	AV	ADV	FF	<i>M</i>	<i>SD</i>
1	1	1	2	-1	-1	-1	-1	-1	28.6	18.5
2	-1	-1	-2	1	1	1	1	1	28.8	19.7
3	1	1	0	2	0	0	0	0	36.1	20.0
4	0	0	0	1	1	1	1	1	39.0	21.8
5	-2	-2	1	1	1	0	0	0	23.0	18.6
6	1	1	1	1	1	2	-2	-2	44.5	20.9
7	1	-2	1	0	0	0	0	0	25.1	18.6
8	0	0	0	0	0	2	2	0	27.6	21.4
9	1	1	1	1	-2	-2	-2	-2	28.9	19.5
10	1	1	1	1	1	-2	0	0	38.5	21.7
11	2	-2	-1	-2	2	0	2	1	33.2	18.5
12	-2	2	2	2	2	2	-2	1	37.6	22.4
13	-2	0	0	-2	-1	-1	-1	2	12.9	16.6
14	0	-1	2	2	2	1	1	0	46.3	22.9
15	2	2	-1	-1	-1	0	-1	-1	27.8	19.8
16	-1	0	-2	0	-2	-1	-2	2	13.3	17.3
17	-1	-1	-2	-1	-1	-2	-1	-2	7.6	15.4
18	0	0	0	0	0	2	0	2	23.1	20.1
19	2	2	2	2	2	2	2	2	74.6	23.7
20	2	2	-1	-2	-2	1	-2	-1	27.2	19.1
All items									31.2	24.4
All items except 17 and 19									30.1	21.8

NOTE: In the survey, the situations were presented in a random order. Safety message sources: CR = average AV crash rate; NM = average AV near-miss rate; FVR = federal vehicle requirements; FG = federal government official position; SLG = state or local government official position; AV = AV company's official position; ADV = safety advocacy group's official position; FF = friends and family members; M = mean; SD = standard deviation.

Message: -2 = strongly shows AVs are unsafe; -1 = mostly shows AVs are unsafe; 0 = no information; 1 = mostly shows AVs are safe; 2 = strongly shows AVs are safe.

Quantitative Messages and Qualitative Messages

Three of the safety message sources referred to statements that are quantitative: average AV crash rate, average AV near-miss rate, and meeting federal vehicle requirements. The other five sources made qualitative statements, in which each source provided a qualitative statement pertaining to the source's judgment that AVs are safe. Three items were designed to capture different information scenarios that might reflect real-world situations. As seen in Table A.2, item 1 described a situation in which quantitative statements supported AV safety while qualitative statements did not. Item 2 described the exact opposite situation. Item 4 described a situation in which there was positive qualitative information but an absence of quantitative

information. Using a repeated-measure analysis of variance that compared the three items, we determined that the average ratings significantly differed ($F[2, 4,326] = 317.96, p < 0.001$). Using individual contrast analyses, item 4 was significantly greater than both items 1 ($t = 22.07, p < 0.001$) and 2 ($t = 21.60, p < 0.001$). Items 1 and 2 did not differ from each other ($t = -0.47, p = 0.64$).

Quantitative Messages and Messages from Government

Two of the message sources were data driven in nature: average AV crash rate and average AV near-miss rate. Three messages originated from the government: meeting federal vehicle requirements, federal government official position, and state or local government official position. Positive and negative messages from these two general source categories were loosely contrasted in items 5 and 15 (Table A.2). These items were compared in a separate repeated-measure analysis of variance. Item 15, which included negative messages from government (and advocacy groups and friends and family) but positive messages from the data-driven sources was rated significantly higher than item 5, which included negative messages from data-driven sources but positive messages from government ($F[1, 2,162] = 96.70, p < 0.001$).

Negative Information from AV Company in Opposition to Other Sources

Item 10 describes a scenario in which many other sources are providing messages that AVs are generally safe while AV companies provide information that AVs are not safe. Individuals in the sample were not particularly influenced by the negative message from AV companies. When analyzed with one-sample t -tests, this item was rated higher on average than the overall mean rating both including ($t(2167) = 13.80, p < 0.001$) and not including ($t(2167) = 16.11, p < 0.001$) the all-negative and all-positive scenarios (items 17 and 19, respectively).

Positive Information from AV Companies with Support from Nongovernment Qualitative Sources

Items 8 and 18 described scenarios in which AVs provided positive safety messages supported by positive messages from either advocacy groups or friends and family, respectively (Table A.2). Comparing the two items, supporting messages originating from advocacy groups significantly improved perceived safety ratings compared with those originating from friends and family ($F[1, 2,163] = 145.07, p < 0.001$).

New Measures Compared with Old Measures

Item 7 describes a scenario in which average AV crash rates and federal vehicle regulations gave information that AVs are safe while AV near-miss rates, a new measure, gave strong information that AVs are unsafe. This pattern of information seemed harmful to perceptions of safety because the perceived safety ratings for this item were on average lower than the overall mean rating both including ($t(2163) = -16.26, p < 0.001$) and not including ($t(2167) = -13.54, p < 0.001$) the all-negative and all-positive scenarios (items 17 and 19, respectively).

Appendix B. Interviews

Semistructured, one-hour interviews were conducted by teleconference with a diverse group of 30 stakeholders during March and April 2020. The interview protocol was piloted with RAND colleagues and revised based on their feedback. Subjects included individuals interviewed for the predecessor project and others. Some individuals invited one or more colleagues to join them. Subjects were promised that what they said would not be attributed to themselves or their organization. One individual encouraged attribution, a couple were indifferent on the issue, and a few pointed to material they had published that could be cited. Research team members took notes during the otherwise unrecorded conversations, and the multiple sets of notes made the records more complete.

Stakeholder Sample

Stratified expert (purposive) sampling was used, with a goal of getting a mix of different kinds of stakeholders that (1) balanced perspectives from categories other than developers (government and safety researchers and advocates) and (2) embodied an aspiration to include as many developers as possible because the industry is so varied. Sampling was constrained principally by the reticence of developers to talk about their work in an intensely competitive marketplace, although the team was also unable to secure an interview with anyone from the insurance industry. Some individuals never responded to invitations (including follow-up messages), and a few declined to participate. The sample was richer than a simple mapping of entity type might suggest because of how senior individuals move professionally in this dynamic arena. Our research benefited from the opportunity to speak with individuals who had experience in both government and industry and who could and did draw on those different experiences explicitly. Because of the small numbers of interviews and the general aversion to attribution, labeling of subjects and quotations has been avoided.

Interview participants were recruited by email, derived from contacts from earlier research by the project team, accumulated professional contacts, suggestions from interviewees, and searches of media coverage. In almost all cases, people who responded to the outreach agreed to be interviewed; some people never responded.

As noted in the acknowledgments, the team interviewed executives or officials from the American Association of State Highway and Transportation Organizations (AASHTO), Advocates for Highway and Automotive Safety, Aurora, Autonocast, the Arizona Department of Transportation, the California Department of Transportation, Consumer Reports, Edge Case Research, the Insurance Institute for Highway Safety (IIHS), Intel, KPMG, Metamoto, Motional, Motus Ventures, the National Institute of Standards and Technology (NIST), the National Safety

Council (NSC), the National Transportation Safety Board (NTSB), NVIDIA, the Pennsylvania Department of Transportation, the San Francisco Metropolitan Transportation Authority, SAE International, the Toyota Research Institute, the Uber Advanced Technologies Group, Voyage, Waymo, and Zoox, as well as transportation expert Jane Lappin, legal scholar Bryant Walker Smith, and a European technologist who preferred to remain anonymous. Describing the distribution of this set must balance the snapshot in time covering the research and the recognition that many of the individuals consulted have had careers that involve moving across organizational categories—a quality that enriched the discussion. In broad strokes, the set can be described as including AV technology developers (10–13), safety researchers and advocates (4–8), AV industry and technology analysts (3–6), and government entities (7–8).

Interview and Note-Taking Guide

The first seven items in the following list represent the interview prompts. The remaining items (8–15) served to bin additional content by category.

1. We are defining “acceptably safe” (and “safe enough”) as being ready for commercial deployment. Do you have a different definition?
2. What qualities should any definition of “acceptably safe” (and “safe enough”) have?
3. Given the AV industry’s diversity, what are the strengths and weaknesses of a uniform method of establishing “acceptably safe” (and “safe enough”) versus each company deciding on its own method?
4. How do public views of “safe enough” factor in?
5. In what ways do people talk about “acceptably safe” (and “safe enough”) that are helpful, and what ways are unhelpful?
6. How does a high level of automation make “acceptably safe” (and “safe enough”) harder? How much of the difficulty is because automated vehicles are a new concept versus other factors?
7. A deeper dive:

In your opinion or experience, what forms of evidence would be the most compelling to you to determine if AVs are safe enough—statistical evidence, regulatory and industry standards, or expert opinions?

Walk through the decision tree [not presented here] [exploring views of different kinds of evidence]

- First selection. Why is that the most convincing for you?
 - Second selection. Why is that the most convincing for you?
 - Third selection. Why is that the most convincing for you?
8. Measure-specific comments not elsewhere classified
 9. Communication-related comments not elsewhere classified
 10. Roadmanship-related comments

11. Cost-benefit and other trade-offs
12. Thoughts on areas of most agreement or disagreement
13. Others to consult
14. Acknowledgment of participation
15. Free-text capture of comments for later binning.

Interview Coding

Qualitative coding of interview notes was completed using the web-based application Dedoose.¹⁸⁷ Interviews were conducted primarily by a lead author of this report, and notes were taken by between two and four other authors. Coding of the notes was completed by three authors. Interviews were semistructured and conducted with stakeholders from diverse backgrounds and expertise within the area of AVs. Therefore, interviews covered a broad variety of topics, necessitating an extensive coding scheme to categorize the collected information. Coding was completed using a set of 150 possible independent codes, categorized under three broad headings: concept of “acceptably safe,” approaches, and direct comparison with other industries. Individual codes were further organized into subcategories within each of those three headings, at one point extending to seven sublevels. Because the interviews were semistructured, the goal of interview coding was to organize the notes into a searchable format rather than for quantitative analysis.

¹⁸⁷ Dedoose Version 8.3.17, web application for managing, analyzing, and presenting qualitative and mixed-method research data, Los Angeles, Calif.: SocioCultural Research Consultants, 2020.

Appendix C. Literature Highlights

Prior to interviewing stakeholders, the research team performed a literature review centered on concepts related to safety that are relevant to AVs. This search included scholarly articles, government reports and other documents, material from standards-setting organizations, material from developers and from industry analysts, and news media. This appendix provides a brief summary of some of the most valuable material we referenced.

General Risk and Safety

Risk, safety, and uncertainty are complex, interconnected topics with definitions that differ depending on their disciplinary field and relative context. Generally, something is “safe” if it is unlikely to cause risk or injury; a thing could be considered safe if it is protected from risk or injury. Nothing is completely safe—there is always a background level of danger, even if that potential danger is very unlikely. Rather than objective endpoints, risk and safety tend to be used as opposing relative terms. Some activities, for example, are “safer” or “riskier” than others. Although there are arguments that risk and safety should not be considered along the same continuum,¹⁸⁸ it is helpful to consider risk reduction when determining what level of safety should be achieved around a potential hazard. Risk can be understood as a function of the probability of the hazard occurring multiplied by its possible severity or harm. Safety from this view is reducing the probability of the hazard or the severity of the hazard to as low as reasonably practical (ALARP) or as low as reasonably achievable (ALARA).¹⁸⁹

The methodologies and studies of risk and safety are used in a wide variety of fields in different yet related ways. Here, we prioritize three of the fields that are most relevant for AVs: engineering, the social sciences, and economics.

In engineering, the emphasis is on risk reduction through identifying “objective” risk—that is, risk that can be quantified with probabilities or expected values.¹⁹⁰ Engineering also uses such methods as Bayesian inference and cost-benefit analysis to identify and mitigate risk while optimizing the outcome.¹⁹¹

The social science approach to risk and safety is more behavioral and focuses on perceptions of risk. Many models within this space focus on perceived costs and benefits, controllability,

¹⁸⁸ Möller, Hansson, and Peterson, 2006.

¹⁸⁹ Rae, 2007; Schäbe, 2001.

¹⁹⁰ Rudolph Frederick Stapelberg, “Safety and Risk in Engineering Design,” in *Handbook of Reliability, Availability, Maintainability and Safety in Engineering Design*, London: Springer, 2009.

¹⁹¹ Terje Aven and Vidar Kristensen, “Perspectives on Risk: Review and Discussion of the Basis for Establishing a Unified and Holistic Approach,” *Reliability Engineering & System Safety*, Vol. 90, No. 1, 2005.

knowledge, and attitudes to determine what risks are “acceptable.”¹⁹² This approach highlights the human elements of risk, including individual judgments, culture (particularly in anthropological contexts),¹⁹³ and the social constructions of risk.¹⁹⁴

Economics focuses on uncertainty, rational decisionmaking, and maximizing expected utility. Risk is captured in probabilities and costs, including the value of a statistical life and other abstractions used to characterize uncertainties and expectations to determine a net present value in myriad variations on cost-benefit analyses.¹⁹⁵ More recently, behavioral economics has sought to understand why individuals make decisions and how those decisions deviate from rational behavior under uncertainty.¹⁹⁶ Both the social science and the economic approaches around safety result not in single solutions but in a variety of solutions based on trade-offs. Consumers require transparent and comprehensive information to compare relative choices. Depending on how risk-averse a consumer is, he or she might prioritize safety, price, convenience, and other personal factors in making decisions about safety.¹⁹⁷

With each of these disciplinary approaches, which should be used to characterize risk for a given hazard? In 2001, the World Health Organization described several possible characterizations about safety levels that are used as a basis for acceptable levels of risk for water-related infectious diseases.¹⁹⁸ These characterizations or approaches provided the framework for our investigation of acceptably safe levels for AVs. The approaches specify that a risk can be considered acceptable when

- it falls below an arbitrary defined probability
- it falls below some level that is already tolerated
- the cost of reducing the risk would exceed the costs saved
- the cost of reducing the risk would exceed the costs saved when the “costs of suffering” are also factored in
- the opportunity costs would be better spent on other, more pressing, problems
- experts or professionals say it is acceptable
- the general public says it is acceptable (or more likely, does not say it is not)
- politicians say it is acceptable.

¹⁹² Fischhoff et al., 1978.

¹⁹³ Sigve Olstedal, Bjorn-Elin Moen, Hroar Klempe, and Torbjørn Rundmo, “Explaining Risk Perception: An Evaluation of Cultural Theory,” *Trondheim: Norwegian University of Science and Technology*, Vol. 85, 2004.

¹⁹⁴ Dorothy Nelkin, “Communicating Technological Risk: The Social Construction of Risk Perception,” *Annual Review of Public Health*, Vol. 10, No. 1, 1989.

¹⁹⁵ William D. Schulze, “Ethics, Economics and the Value of Safety,” in *Societal Risk Assessment*, Boston, Mass.: Springer, 1980.

¹⁹⁶ Lisa A. Robinson and James K. Hammitt, “Behavioral Economics and Regulatory Analysis,” *Risk Analysis: An International Journal*, Vol. 31, No. 9, 2011.

¹⁹⁷ Andreas Oehler and Stefan Wendt, “Good Consumer Information: The Information Paradigm at Its (Dead) End?” *Journal of Consumer Policy*, Vol. 40, No. 2, 2017.

¹⁹⁸ Fewtrell and Bartram, 2001.

Technology Acceptance and Risk Perception

The study of risk perception in psychology was greatly inspired by initial investigations into the perception of risks associated with new technologies, particularly with the expansion of nuclear power plants in the United States in the 1970s and the associated concern from the lay public. In general, technologies that are perceived to be beneficial and to be low cost (e.g., financial, learning costs, opportunity costs) promote technology acceptance; this, in turn, influences technology adoption and usage intention, which, again in turn, promote actual usage behavior. Covered more thoroughly in Chapter 6, the works listed here highlight some of the broader theoretical accounts of technology adoption and specific studies that are relevant to human perception of automation, transportation technologies, and other novel technologies with elements of risk.

Risk Perception and Technology and Transportation

Ewart J. de Visser, Richard Pak, and Tyler H. Shaw, “From ‘Automation’ to ‘Autonomy’: The Importance of Trust Repair in Human-Machine Interaction,” *Ergonomics*, Vol. 61, No. 10, 2018

Authors use human-centered approaches to promote human trust of machines, including an industrial-psychology-informed framework for automated systems and a model for guidance in future automated systems research.

Peng Liu, Yong Du, and Zhigang Xu, “Machines Versus Humans: People’s Biased Responses to Traffic Accidents Involving Self-Driving Vehicles,” *Accident Analysis & Prevention*, Vol. 125, 2019

Using vignette-based experiments, researchers found that participants tended to perceive crashes involving self-driving vehicles as more severe than those with only human drivers—regardless of whether the automated system was responsible and regardless of any injuries sustained.

John M. Osepchuck, “The History of the Microwave Oven: A Critical Review,” *2009 IEEE MMT-S International Microwave Symposium Digest*, 2009

This brief account of the history and context of the production and public rollout of microwave ovens includes public perceptions of associated risk.

Md Mahmudur Rahman, Mary F. Lesch, William J. Horrey, and Lesley Strawderman, “Assessing the Utility of TAM, TPB, and UTAUT for Advanced Driver Assistance Systems,” *Accident Analysis & Prevention*, Vol. 108, 2017

Researchers modeled driver acceptance of ADAS using three different models: Technology Acceptance Model (TAM), the Theory of Planned Behavior (TPB), and the Unified Theory of Acceptance and Use of Technology (UTAUT). Using a survey approach,

they found that each of these models individually explained 71 percent or more of the variability in behavioral intention, with the TAM performing the best.

Torbjørn Rundmo, Trond Nordfjærn, Hilde Hestad Iversen, Sigve Olteidal, and Stig H. Jørgensen, “The Role of Risk Perception and Other Risk-Related Judgements in Transportation Mode Use,” *Safety Science*, Vol. 49, No. 2, 2011

Using a survey, the authors found that perceived control of modes of transportation, knowledge about safety, and trust in authority were significantly different between individuals who used private modes of transportation and those who used public modes.

Chauncey Starr, “Social Benefit Versus Technological Risk,” *Science*, Vol. 165, No. 3899, 1969

This early and seminal work on “how safe is safe enough” described a quantitative measure of benefits relative to costs based on risks of disruptive technologies used by the public. Starr found that the public is willing to accept “voluntary” risks roughly 1,000 times greater than “involuntary” risks and that the social acceptance of risk is affected by the public’s awareness of the benefits and usefulness of a technology.

Models of Technology Acceptance

Automation Acceptance Model

Mahtab Ghazizadeh, John D. Lee, and Linda Ng Boyle, “Extending the Technology Acceptance Model to Assess Automation,” *Cognition, Technology & Work*, Vol. 14, No. 1, 2012

The authors use this extension of the TAM to describe how trust and the compatibility between task and technology can influence perceptions of usefulness, ease of use, and technology attitudes, which in turn influence behavior intention and use of automation technologies.

Technology Acceptance Model

Fred D. Davis, Richard P. Bagozzi, and Paul R. Warshaw, “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models,” *Management Science*, Vol. 35, No. 8, 1989

Using the TAM, authors describe how perceived usefulness and ease of use influences attitudes toward technology, which in turn influence behavioral intention and technology use.

Unified Theory of Acceptance and Use of Technology

Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis, “User Acceptance of Information Technology: Toward a Unified View,” *MIS Quarterly*, Vol. 27, No. 3, 2003

Inspired by the TAM, the authors describe the UTAUT and explain how social influences and perceptions of usefulness and ease of use can influence behavioral intentions, which, along with facilitating conditions, influence technology use.

Viswanath Venkatesh, James Y. L. Thong, and Xin Xu, “Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology,” *MIS Quarterly*, Vol. 36, No. 1, 2012.

The authors create an extension of the UTAUT model, introducing the additional factors of hedonic motivation, price value, and habit as predictors of behavioral intention and, for the habit factor, technology use.

Perceptions of AVs

Among studies of technology attitudes, survey studies on the perception of AVs are comparatively short-lived: The industry is constantly in flux, and public perception is fluid in response. Although change over time is found in many studies of public perception in other domains, change in perceptions related to AVs is particularly dynamic considering the rapid advancement in the field and the public’s general lack of experience with the technology. Therefore, studies on perceptions of AVs from even a few years ago can be considered out of date. This section highlights several relatively recent survey studies investigating the perceptions of AV, and some older studies that include investigations of specific population groups.

General Surveys on AV Perceptions

Neil Charness, Jong Sung Yoon, Dustin Souders, Cary Stothart, and Courtney Yehnert, “Predictors of Attitudes Toward Autonomous Vehicles: The Roles of Age, Gender, Prior Knowledge, and Personality,” *Frontiers in Psychology*, Vol. 9, 2018

Using a sample of Amazon Mechanical Turk middle-aged crowdworkers, this study found three factors that affected user attitudes: concern with AV, eagerness to adopt AV technology, and willingness to relinquish driving control. Other individual aspects, including age, gender, prior knowledge, and personality traits, also were found to modify attitudes.

Nikhil Menon, Yu Zhang, Abdul Rawoof Pinjari, and Fred Mannering, “A Statistical Analysis of Consumer Perceptions Towards Automated Vehicles and Their Intended Adoption,” *Transportation Planning and Technology*, Vol. 43, No. 3, 2020

Using cluster analysis of survey data, the authors investigated consumer perceptions of benefits and concerns regarding AVs and their likelihood of adoption. They identified four “market segments” of interest: benefits-dominated, concerns-dominated, uncertain, and

well-informed. The authors also identify the possibility of large gaps in AV adoption between generational and demographic groups.

Joanna Moody, Nathaniel Bailey, and Jinhua Zhao, “Public Perceptions of Autonomous Vehicle Safety: An International Comparison,” *Safety Science*, Vol. 121, 2020

Using a large international sample and structural equation modeling, researchers identified differences in perceptions about AV safety and deployment. They found that young, employed, high-income, and highly educated males are the most optimistic about AV safety, and they found that people in developing countries in Asia are also optimistic about AV safety.

Ipek N. Sener, Johanna Zmud, and Chris Simek, *Examining Future Automated Vehicle Usage: A Focus on the Role of Ride Hailing*, College Station, Tex.: Texas A&M Transportation Institute, May 2018

Using an online survey and interviews, the authors compared the likelihood of acceptance and use of self-driving vehicles among individuals who use ride-hailing services (such as Uber and Lyft) with those who do not use such services. The authors found that the intent to use AV was “exceptionally higher” among users of ride-hailing services—particularly long-term users.

Qiaoning Zhang, X. Jessie Yang, and Lionel Peter Robert, “Expectations and Trust in Automated Vehicles,” *CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, April 2020

Researchers performed a polynomial regression on the results of an online survey to investigate trust in AVs using the expectation-confirmation theory. They found that trust is higher when perceived performance is higher than individuals’ expectations and “perceived risk can moderate the relationship between expectation confirmation and trust in AVs.”

Specific Population Groups

Rosaria M. Berliner, Scott Hardman, and Gil Tal, “Uncovering Early Adopter’s Perceptions and Purchase Intentions of Automated Vehicles: Insights from Early Adopters of Electric Vehicles in California,” *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 60, January 2019

Drawing on a sample of new electrical vehicle buyers in California, the authors found that young men who bought high-cost vehicles were most interested in obtaining AVs, about one-half of the respondents had average knowledge of AVs, and those who perceived AVs to be safer than non-AVs were most likely to be interested in acquiring one.

Julian Brinkley, Juan E. Gilbert, and Shaundra B. Daily, “A Survey of Visually Impaired Consumers About Self-Driving Vehicles,” in *33rd Annual International Technology and Persons with Disabilities Conference Scientific/Research Proceedings*, San Diego, 2018

Researchers surveyed individuals over the age of 18 who identified as visually impaired and found that, although respondents had a generally positive impression of the technology, there were concerns about equipment failure, unexpected situations that the AV could not handle, and interactions between the AV and pedestrians or cyclists.

David W. Eby, Lisa J. Molnar, and Sergiu C. Stanciu, *Older Adults' Attitudes and Opinions About Automated Vehicles: A Literature Review*, Ann Arbor, Mich., and College Station, Tex.: ATLAS Center, ATLAS-2018-26, 2018

The authors reviewed ten studies focused on the perceptions of older adults concerning AVs. Results indicated that the majority of older adults were reluctant to use vehicles that are fully automated, that they did not intend to own any, and that they had neutral to negative opinions of AV technology.

Peng Liu, Zhang Yawen, and He Zhen, "The Effect of Population Age on the Acceptability of Safety of Self-Driving Vehicles," *Reliability Engineering & System Safety*, Vol. 185, Iss. C, 2019

Researchers sought to identify age differences in AV acceptability from using a sample of residents of Tianjin, China. They found that younger respondents had more-positive attitudes and acceptance of AV compared with older respondents and that the older population are likely to prefer higher standards of safety for AVs.

Anton Manfreda, Klara Ljubi, and Aleš Groznik, "Autonomous Vehicles in the Smart City Era: An Empirical Study of Adoption Factors Important for Millennials," *International Journal of Information Management*, December 2019

The researchers surveyed millennials for their perceptions of AVs and found that the sample was primarily interested in the perceived benefits of AVs; the perceived safety of the technology was seen more negatively.

Ericka Rovira, Anne Collins McLaughlin, Richard Pak, and Luke High, "Looking for Age Differences in Self-Driving Vehicles: Examining the Effects of Automation Reliability, Driving Risk, and Physical Impairment on Trust," *Frontiers in Psychology*, Vol. 10, 2019

Using a survey of older and younger adults, the authors sought to identify age-based differences in trust of AVs. They found that multiple interacting variables had an influence on AV trust, including the age of the respondent, travel risk, impairment level of a hypothesized driver, and whether the self-driving car was reliable.

Carley Ward, Martina Raue, Chaiwoo Lee, Lisa D-Ambrosio, and Joseph F. Coughlin, "Acceptance of Automated Driving Across Generations: The Role of Risk and Benefit Perception, Knowledge, and Trust," in Masaaki Kurosu, ed., *Human-Computer Interaction:*

User Interface Design, Development, and Multimodality 19th International Conference, HCI International 2017 Proceeding, Part I, New York: Springer International Publishing, 2017

A survey of U.S. adults was assessed to explore how generational age, knowledge, and trust affect individual perceptions of AV acceptability. The results corroborated other studies that found that trust, knowledge, and perceptions of risk and benefit are important for AV acceptance and that respondents' attitudes about AVs were dependent on age group and gender. The authors found that informational materials have the potential to boost positive feelings and the perceived benefit of AVs.

Safety Culture

Safety culture describes a general set of beliefs, values, and norms related to safety that are held at the organizational level. It has been a topic of interest across numerous industries and academic disciplines. The study of safety culture has produced several theoretical models and questionnaires designed to measure safety climate, a snapshot measurement of the product of safety culture on organizational values at a particular moment in time. Researchers examined safety culture and climate and the psychological constructs of personality and mood. In this case, *safety culture* is akin to the more stable trait of personality, while *safety climate* is akin to the more transient state of mood.¹⁹⁹ Because of its quantitative nature and focus on measurement, the state of research on safety climate is more developed than that of safety culture and has produced several survey instruments, though many are domain- or context-specific. AV development is a relatively new industry, so an AV-specific safety climate measure does not yet exist, to our knowledge. Although this might seem to be a barrier for application within the AV industry, a meta-analytic review provided evidence that general measures were more predictive of some safety outcomes (e.g., errors, near misses, noninjury safety-related events) than industry specific measures.²⁰⁰

Safety Culture: Select Integrative Models

Tiffany M. Bisbey, Molly P. Kilcullen, Eric J. Thomas, Madelene J. Ottosen, KuoJen Tsao, and Eduardo Salas, "Safety Culture: An Integration of Existing Models and a Framework for Understanding Its Development," *Human Factors*, August 19, 2019

¹⁹⁹ Sue J. Cox and Rhona Flin, "Safety Culture: Philosopher's Stone or Man of Straw?" *Work & Stress*, Vol. 12, No. 3, 1998.

²⁰⁰ Lixin Jiang, Lindsey M. Lavaysse, and Tahira M. Probst, "Safety Climate and Safety Outcomes: A Meta-Analytic Comparison of Universal vs. Industry-Specific Safety Climate Predictive Validity," *Work & Stress*, Vol. 33, No. 1, 2019.

The authors present an integration of existing safety culture models, suggesting that safety culture is developed through a bottom-up process (e.g., employees learning from safety outcomes) and facilitated by various enabling factors.

M. Dominic Cooper, “Towards a Model of Safety Culture,” *Safety Science*, Vol. 36, 2000

This article proposes a measurement model for safety culture, integrating reciprocal factors of safety attitudes and perceptions, safety behavior, and context (e.g., an institutional safety management system).

Geert Vierendeels, Genserik Reniers, Karolien van Nunen, and Koen Ponnet, “An Integrative Conceptual Framework for Safety Culture: The Egg Aggregated Model (TEAM) of Safety Culture,” *Safety Science*, Vol. 103, 2018

This article introduces “The Egg Aggregated Model” (TEAM), a broad conceptual framework for safety culture that describes a cyclical relationship between safety perceptions, behavioral intentions, and observable safety outcomes.

Safety Climate Measures: Selected Review and Domain-General Survey Instruments

Susan E. Hahn and Lawrence R. Murphy, “A Short Scale for Measuring Safety Climate,” *Safety Science*, Vol. 46, 2008

This six-item safety climate scale provides a measure with general items adaptable to specific contexts.

Lixin Jiang, Lindsey M. Lavaysse, and Tahira M. Probst, “Safety Climate and Safety Outcomes: A Meta-Analytic Comparison of Universal vs. Industry-Specific Safety Climate Predictive Validity,” *Work & Stress*, Vol. 33, No. 1, 2019

This article provides a meta-analytic review of predictive capabilities of safety climate measures. Universal safety climate measures have more predictive power for non-injury- or non-accident-related adverse events, although industry-specific safety climate measures are better predictors of safety behaviors and risk perceptions.

Pete Kines, Jorma Lappalainen, Kim Lyngby Mikkelsen, Espen Olsen, Anders Pousette, Jorunn Tharaldsen, Kristinn Tómasson, and Marianne Törnerd, “Nordic Safety Climate Questionnaire (NOSACQ-50): A New Tool for Diagnosing Occupational Safety Climate,” *International Journal of Industrial Ergonomics*, Vol. 41, No. 6, November 2011.

This 50-item Nordic Safety Climate Questionnaire (NOSACQ-50) is a long-form context-free survey instrument.

Selected Industry- and Government-Recommended Standards and Guidelines

During the literature scan, and throughout the interviews, many existing documents from industry, U.S. government, and nonprofit bodies were raised as examples of standards that are shaping, or could shape, the AV industry with regard to safety. Here, we provide a selected list of the documents that were most salient during the process.

AVSC, *AVSC Best Practice for Describing an Operational Design Domain: Conceptual Framework and Lexicon*, Warrendale, Pa.: SAE Industry Technologies Consortia, AVSC00002202004, 2020

Drafted by the AVSC, which is a program from the SAE Industry Technologies Consortia, this document seeks to provide common language around the ODDs that AV developers use in testing.

California Department of Motor Vehicles, “Autonomous Vehicles,” webpage, undated and

California Public Utilities Commission “Autonomous Vehicle Passenger Service Pilot Programs,” webpage, undated

California enacted two pilot programs to enable regulated testing within its borders. Before a manufacturer can join the program, it must acquire a permit from the Department of Motor Vehicles in accordance with California Code of Regulations, Title 13, Article 3.7.

Institute of Electrical and Electronics Engineers Standards Association, “P2846—A Formal Model for Safety Considerations in Automated Vehicle Decision Making,” webpage, undated-b

IEEE is developing this standard to create mathematical, rule-based standards that will be applicable to any technology developer and customizable for different areas. Part of the basis of this standard is Intel’s RSS framework.

ISO, “Road Vehicles—Safety of the Intended Functionality,” ISO/PAS 21448:2019, 2019

These standards address hazards from the intended functionality of a system. That is, this standards document provides “guidance on the applicable design, verification, and validation of measures needed to achieve SOTIF.” ISO/PAS 21448 differs from ISO 26262 in that the former covers safety hazards that occur during normal operation while the latter addresses functional safety in the event of a systems failure.

ISO, “Road Vehicles—Functional Safety—Part 1: Vocabulary,” ISO 26262-1:2011, 2011

This standard focuses on the functional safety of electrical and electronic systems, particularly in the case of their malfunction or combination with other systems.

NHTSA, “Regulations,” webpage, undated-c

FMVSS, U.S. federal regulations for all domestic vehicles, have three main purposes: crash avoidance, crashworthiness, and post-crash survivability. These standards are enforced by NHTSA. Rarely, exemptions are granted from FMVSS for a few specific reasons—one of which includes “facilitation of the development of new motor vehicle safety or low-emission engine features, or existence of an equivalent overall level of motor vehicle safety.” Exemption from these standards is one route considered for the development of AVs.

NHTSA, “Voluntary Safety Self-Assessment,” webpage, undated-d

NHTSA defined a set of 12 safety elements that AV developers can voluntarily submit to publicly disclose how they are addressing public safety concerns. Those companies that have submitted voluntary assessments are listed in an online disclosure index.

SAE International, “Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles,” Standard J3016_201806, June 15, 2018

Released from SAE International, a standards-developing organization, this standard provides a taxonomy of automated driving capabilities, ranging from level 0 (no driving automation) to level 5 (fully automated driving with no human needed).

Underwriters Laboratories Standards, “Standard for Evaluation of Autonomous Products,” Standard 4600, Edition 1, April 1, 2020

Underwriters Laboratories developed this standard in conjunction with Edge Case Research to be “goal based and technology-agnostic.” As a performance-based standard, it focuses on creating a safety case with identifiable goals, argumentation about why a goal is achieved, and evidence to support that the arguments are valid.

Safety in Other Industries

Other industries, in both transportation and other domains, have longer research histories that can provide insight into safety-related issues within the AV field. Here, we describe selected reports and research documents that address safety issues within other industries that operate within a spectrum of risk.

Aviation

Allianz Global Corporate & Specialty, *Global Aviation Safety Study: A Review of 60 Years of Improvement in Aviation Safety*, Munich, Germany, December 2014

This retrospective safety analysis from a mature transportation industry summarizes safety challenges, successes, and evolution in the history of the aviation industry.

Federal Aviation Administration, “System Design and Analysis,” Advisory Circular 25.1309-1A, June 21, 1988

This circular documents the steps necessary to demonstrate an inverse relationship between the probability and severity of adverse events. It clarifies regulations for aviation designed to balance risk and the severity of negative outcomes.

Antonius A. Lambregts, “Flight Envelope Protection for Automatic and Augmented Manual Control,” in *Proceedings of the EutoGNC 2013, 2nd CEAS Specialist Conference on Guidance, Navigation and Control*, 2013

The author highlights human factors of transitions between automated and manual control and provides suggestions for Flight Envelope Protection integration between automated and manual operation.

Dave Michaels, Andy Pasztor, and Andrew Tangel, “Two FAA Officials Are Key Witnesses in Criminal Probe of Ex-Boeing Pilot,” *Wall Street Journal*, March 14, 2020

This article describes a potential negative outcome of the reliance of regulatory bodies on industry insiders (or experts) for their understanding of the functionality of cutting-edge technology systems. It expands on the role of a Boeing pilot in misleading regulators in the capabilities of the 737 MAX planes prior to the 2018 and 2019 crashes.

U.S. House of Representatives, Committee on Transportation and Infrastructure, *The Boeing 737 MAX Aircraft: Costs, Consequences, and Lessons from Its Design, Development, and Certification*, final committee report, September 2020

This description of an investigation into a vehicle manufacturer after adverse events resulting in loss of life summarizes takeaways from the preliminary investigation of two 737 MAX crashes in 2018 and 2019.

Biotechnology

Joanna K. Sax and Neal Doran, “Ambiguity and Consumer Perceptions of Risk in Various Areas of Biotechnology,” *Journal of Consumer Policy*, Vol. 42, No. 1, 2019

This study indicates that receiving ambiguous information pertaining to new technologies results in perceptions of higher risk and lower benefits. Ambiguous information and aversion to ambiguity is related to lower perceptions of benefits and higher perceptions of risk in biotechnology (e.g., foods, vaccines, fluoridated water, and stem cell research).

Electric Scooters

Alexandre Santacreu, George Yannis, Omblin de Saint Léon, and Philippe Crist, *Safe Micromobility*, Paris, France: International Transport Forum, 2020

The authors explore integrating a new vehicular technology into an existing transportation environment in this safety report on micromobility technologies.

Genetically Modified Foods

Joachim Scholderer and Lynn J. Frewer, “The Biotechnology Communication Paradox: Experimental Evidence and the Need for a New Strategy,” *Journal of Consumer Policy*, Vol. 26, 2003

The authors investigate the difficulty of convincing the public of the merits of new technologies in the presence of preexisting negative associations. This experimental study shows that overt attempts to convince the public of the safety of genetically modified organisms (GMOs) activates, rather than supersedes, preexisting attitudes.

Jennifer B. Wohl, “Consumer’s Decision-Making and Risk Perceptions Regarding Foods Produced with Biotechnology,” *Journal of Consumer Policy*, Vol. 21, 1998

This study showed that acceptability of genetically modified goods is influenced both by the magnitude and characteristics of the associated risk. The author describes how the quantitative aspects and qualitative nature of risks (see the psychometric paradigm of risk discussed in Chapter 6) influence the acceptability of a new technology.

Global Warming

Consortium for Science, Policy & Outcomes at Arizona State University, *Cooling a Warming Planet? Public Forums on Climate Intervention Research*, Tempe, Ariz., November 2019

This report on a series of focus groups designed to elucidate public attitudes and preferences for research on global warming gauges public attitudes on an advanced technological solution to a naturalistic problem in a politically polarized environment.

Medical Devices

Judith A. Johnson, *FDA Regulation of Medical Devices*, Washington, D.C.: Congressional Research Service, R42130, September 2016

This report summarizes the regulatory strategy for medical devices pre- and post-market. It describes how regulation occurs before and after product launch in a diverse industry and lists potential benefits and associated risks.

Pharmaceuticals

Agata Dabrowska and Susan Thaul, *How FDA Approves Drugs and Regulates Their Safety and Effectiveness*, Washington, D.C.: Congressional Research Service, R41983, May 8, 2018

This report describes how regulation occurs before and after product launch in a diverse industry with potential benefits and associated risks and summarizes the regulatory strategy for medical drugs pre- and post-market.

Bibliography

- Abraham, Hillary, Bobbie Seppelt, Bruce Mehler, and Bryan Reimer, “What’s in a Name: Vehicle Technology Branding & Consumer Expectations for Automation,” *Proceedings of the 9th ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2016, pp. 226–234.
- Ackermans, Sander, Debargha Dey, Peter Ruijten, Raymond H. Cuijpers, and Bastian Pfleging, “The Effects of Explicit Intention Communication, Conspicuous Sensors, and Pedestrian Attitude in Interactions with Automated Vehicles,” *Proceedings of the CHI 2020 Conference on Human Factors in Computing Systems*, Paper 70, 2020, pp. 1–14.
- Allianz Global Corporate & Specialty, *Global Aviation Safety Study: A Review of 60 Years of Improvement in Aviation Safety*, Munich, Germany, December 2014. As of July 20, 2020: <https://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/AGCS-Global-Aviation-Safety-2014-report.pdf>
- American Automobile Association, Consumer Reports, J. D. Power, National Safety Council, and SAE International, “Clearing the Confusion: Recommended Common Naming for Advanced Driver Assistance Technologies,” webpage, undated. As of August 30, 2020: <https://www.sae.org/binaries/content/assets/cm/content/miscellaneous/adas-nomenclature.pdf>
- American National Standards Institute, “Project Initiation Notification System (PINS),” *ANSI Standards Action*, Vol. 51, No. 27, July 30, 2020. As of September 10, 2020: <https://share.ansi.org/Shared%20Documents/Standards%20Action/2020-PDFs/SAV5127.pdf>
- Aptiv Services, Audi, Bayrische Motoren Werke (BMW), Beijing Baidu Netcom Science Technology, Continental Teves, Daimler, Fiat Chrysler Automobiles (FCA), HERE Global B.V., Infineon Technologies, Intel, and Volkswagen, *Safety First for Automated Driving*, 2019. As of August 30, 2020: <https://www.daimler.com/documents/innovation/other/safety-first-for-automated-driving.pdf>
- Arai, Yuji, Tetsuya Nishimoto, Yukihiro Ezaka, and Kenichi Yoshimoto, “Accidents and Near-Misses Analysis by Using Video Drive-Recorders in a Fleet Test,” in *Proceedings of the 17th International Technical Conference on the Enhanced Safety of Vehicles (ESV) Conference*, Amsterdam, 2001.
- Automated Vehicle Safety Consortium, *AVSC Best Practice for Describing an Operational Design Domain: Conceptual Framework and Lexicon*, Warrendale, Pa.: SAE Industry Technologies Consortia, AVSC00002202004, 2020. As of August 31, 2020: <https://www.sae.org/standards/content/avsc00002202004/preview/>

- Automotive IQ, “Car Safety: History and Requirements of ISO 26262,” webpage, June 29, 2016. As of June 28, 2020:
<https://www.automotive-iq.com/electrics-electronics/articles/the-history-and-requirements-of-iso-26262>
- Aven, Terje, “On the Ethical Justification for the Use of Risk Acceptance Criteria,” *Risk Analysis*, Vol. 27, No. 2, April 2007, pp. 303–312.
- Aven, Terje, and Vidar Kristensen, “Perspectives on Risk: Review and Discussion of the Basis for Establishing a Unified and Holistic Approach,” *Reliability Engineering & System Safety*, Vol. 90, No. 1, 2005, pp. 1–14.
- AVSC—See Automated Vehicle Safety Consortium.
- Azevedo-Sa, Herbert, Suresh Kumaar Jayaraman, Connor T. Esterwood, X. Jessie Yang, Lionel P. Robert Jr., and Dawn M. Tilbury, “Comparing the Effects of False Alarms and Misses on Human’s Trust in (Semi)Autonomous Vehicles,” *HRI ’20 Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, New York: Association for Computing Machinery, 2020, pp. 113–115.
- Barickman, Frank, Joshua L. Every, Bowen Weng, Scott Schnelle, and Sughosh Rao, “Instantaneous Safety Metric,” National Highway Traffic Safety Administration, PowerPoint presentation, June 25, 2019. As of August 28, 2020:
https://www.nist.gov/system/files/documents/2019/07/12/day1_part3_barickman_nist_ism_presentation.pdf
- Beggiato, Matthias, and Josef F. Krems, “The Evolution of Mental Model, Trust and Acceptance of Adaptive Cruise Control in Relation to Initial Information,” *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 18, 2013, pp. 47–57.
- Berliner, Rosaria M., Scott Hardman, and Gil Tal, “Uncovering Early Adopter’s Perceptions and Purchase Intentions of Automated Vehicles: Insights from Early Adopters of Electric Vehicles in California,” *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 60, January 2019, pp. 712–722.
- Bin-Nun, Amitai Y., Anthony Panasci, and Radboud J. Duintjer Tebbens, *Heinrich’s Triangle, Heavy-Tailed Distributions, and Autonomous Vehicle Safety*, presented at the 99th Transportation Research Board annual meeting, Washington, D.C., January 2020.
- Bisbey, Tiffany M., Molly P. Kilcullen, Eric J. Thomas, Madelene J. Ottosen, KuoJen Tsao, and Eduardo Salas, “Safety Culture: An Integration of Existing Models and a Framework for Understanding Its Development,” *Human Factors*, August 19, 2019.
- Bloomfield, Robin E., and Peter Bishop, “Safety and Assurance Cases: Past, Present and Possible Future—an Adelard Perspective,” in Chris Dale and Tom Anderson, eds., *Making*

Systems Safer: Proceedings of the Eighteenth Safety-Critical Systems Symposium, Bristol, UK, 9-11th February 2010, London: Springer, 2010, pp. 51–67.

Brinkley, Julian, Juan E. Gilbert, and Shaundra B. Daily, “A Survey of Visually Impaired Consumers About Self-Driving Vehicles,” in *33rd Annual International Technology and Persons with Disabilities Conference Scientific/Research Proceedings*, San Diego, 2018.

British Standards Institution, “Assuring the Safety of Automated Vehicle Trials and Testing—Specification,” Publicly Available Specification 1881, London, 2020. As of July 20, 2002: <https://www.bsigroup.com/en-GB/CAV/pas-1881/>

California Department of Motor Vehicles, “Autonomous Vehicles,” webpage, undated. As of July 20, 2020: <https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/>

California Public Utilities Commission, “Autonomous Vehicle Passenger Service Pilot Programs,” webpage, undated. As of August 30, 2020: <https://www.cpuc.ca.gov/avcpilotinfo/>

Campbell, Kenneth L., “The SHRP 2 Naturalistic Driving Study: Addressing Driver Performance and Behavior in Traffic Safety,” *TR News*, No. 282, September–October 2012, pp. 30–36. As of August 31, 2020: https://insight.shrp2nds.us/documents/shrp2_background.pdf

Cassani Davis, Lauren, “Would You Pull the Trolley Switch? Does It Matter? The Lifespan of a Thought Experiment,” *The Atlantic*, October 2015. As of August 28, 2020: <https://www.theatlantic.com/technology/archive/2015/10/trolley-problem-history-psychology-morality-driverless-cars/409732/>

Censi, Andrea, Konstantin Slutsky, Tichakorn Wongpiromsarn, Dmitry Yershov, Scott Pendleton, James Fu, and Emilio Frazzoli, *Liability, Ethics, and Culture-Aware Behavior Specification Using Rulebooks*, Dublin, Ireland: Aptiv, white paper, 2019. As of June 28, 2020: <https://www.aptiv.com/docs/default-source/white-papers/aptiv-rulebooks.pdf>

Centers for Disease Control and Prevention, “Teen Drivers,” webpage, October 28, 2019. As of August 20, 2020: https://www.cdc.gov/motorvehiclesafety/teen_drivers/index.html

Chandraratna, Susantha, Nick Stamatiadis, and Arnold Stromberg, “Crash Involvement of Drivers with Multiple Crashes,” *Accident Analysis & Prevention*, Vol. 38, No. 3, June 2006, pp. 532–541. As of August 30, 2020: https://www.researchgate.net/publication/7365874_Crash_involvement_of_drivers_with_multiple_crashes

- Chao, Elaine L., U.S. Department of Transportation Secretary, remarks as prepared for delivery at the Autonomous Vehicle Symposium, San Francisco, Calif., July 10, 2018. As of July 20, 2020:
<https://www.transportation.gov/briefing-room/7-10-2018-autonomous-vehicle-symposium>
- Charness, Neil, Jong Sung Yoon, Dustin Souders, Cary Stothart, and Courtney Yehnert, “Predictors of Attitudes Toward Autonomous Vehicles: The Roles of Age, Gender, Prior Knowledge, and Personality,” *Frontiers in Psychology*, Vol. 9, 2018, pp. 1–9.
- Christmann, Petra, and Glen Taylor, *Firm Self-Regulation Through International Certifiable Standards: Determinants of Symbolic Versus Substantive Implementation*, paper submitted to the first annual Conference on Institutional Mechanisms for Industry Self-Regulation, Dartmouth University, Hanover, N.H., November 14, 2005. As of August 30, 2020:
<http://mba.tuck.dartmouth.edu/mechanisms/pages/Papers/ChristmannTaylorDartmouth.pdf>
- Consortium for Science, Policy & Outcomes at Arizona State University, *Cooling a Warming Planet? Public Forums on Climate Intervention Research*, Tempe, Ariz., November 2019. As of August 30, 2020:
<https://cspo.org/publication/srmfinalreport/>
- Cooksey, Ray W., “The Methodology of Social Judgement Theory,” *Thinking & Reasoning*, Vol. 2, No. 2–3, 1996, pp. 141–174.
- Cooper, M. Dominic, “Towards a Model of Safety Culture,” *Safety Science*, Vol. 36, 2000, pp. 111–136.
- Cox, Sue J., and Rhona Flin, “Safety Culture: Philosopher’s Stone or Man of Straw?” *Work & Stress*, Vol. 12, No. 3, 1998, pp. 189–201.
- Dabrowska, Agata, and Susan Thaul, *How FDA Approves Drugs and Regulates Their Safety and Effectiveness*, Washington, D.C.: Congressional Research Service, R41983, May 8, 2018. As of July 20, 2020:
<https://fas.org/sgp/crs/misc/R41983.pdf>
- Das, Subasish, Xiaoduan Sun, Fan Wang, and Charles Leboeuf, “Estimating Likelihood of Future Crashes for Crash-Prone Drivers,” *Journal of Traffic and Transportation Engineering*, Vol. 2, No. 3, June 2015, pp. 145–147. As of August 31, 2020:
<https://www.sciencedirect.com/science/article/pii/S2095756415000252>
- Davis, Fred D., Richard P. Bagozzi, and Paul R. Warshaw, “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models,” *Management Science*, Vol. 35, No. 8, 1989, pp. 982–1003.

- de Visser, Ewart J., Richard Pak, and Tyler H. Shaw, “From ‘Automation’ to ‘Autonomy’: The Importance of Trust Repair in Human-Machine Interaction,” *Ergonomics*, Vol. 61, No. 10, 2018, pp. 1409–1427.
- Dedoose Version 8.3.17, web application for managing, analyzing, and presenting qualitative and mixed-method research data, Los Angeles, Calif.: SocioCultural Research Consultants, 2020. As of June 23, 2020:
<https://app.dedoose.com/App/?Version=8.3.17>
- Denney, Ewen, Ganesh Pai, and Josef Pohl, “Heterogeneous Aviation Safety Cases: Integrating the Formal and the Non-Formal,” in *IEEE 17th International Conference on Engineering of Complex Computer Systems*, Paris: Institute of Electrical and Electronics Engineers, 2012, pp. 199–208. As of June 29, 2020:
<https://ieeexplore.ieee.org/document/6299215>
- Du, Na, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K. Pradhan, X. Jessie Yang, and Lionel P. Robert Jr., “Look Who’s Talking Now: Implications of AV’s Explanations on Driver’s Trust, AV Preference, Anxiety, and Mental Workload,” *Transportation Research Part C: Emerging Technologies*, Vol. 104, July 2019, pp. 428–442.
- Duncan, Ian, “EasyMile Autonomous Shuttles Barred from Carrying Passengers,” *Washington Post*, February 29, 2020.
- Eby, David W., Lisa J. Molnar, and Sergiu C. Stanciu, *Older Adults’ Attitudes and Opinions About Automated Vehicles: A Literature Review*, Ann Arbor, Mich., and College Station, Tex.: ATLAS Center, ATLAS-2018-26, 2018.
- Ecola, Liisa, Steven W. Popper, Richard Silberglitt, and Laura Fraade-Blanar, *The Road to Zero: A Vision for Achieving Zero Roadway Deaths by 2050*, National Safety Council and RAND Corporation, RR-2333-NSC, 2018. As of June 26, 2020:
https://www.rand.org/pubs/research_reports/RR2333.html
- Eliot, Lance, “Essential Stats for Justifying and Comparing Self-Driving Cars to Humans at the Wheel,” *Forbes*, May 30, 2019. As of June 28, 2020:
<https://www.forbes.com/sites/lanceeliot/2019/05/30/essential-stats-for-justifying-and-comparing-self-driving-cars-to-humans-at-the-wheel/#48ad8f9446ed>
- Engström, Johan, Andrew Miller, Wenyan Huang, Susan Soccolich, Sahar Ghanipoor Machiani, Arash Jahangiri, Felix Dreger, and Joost de Winter, *Behavior-Based Predictive Safety Analytics—Pilot Study*, Blacksburg, Va.: Virginia Tech Transportation Institute, April 2019.
- Every, Joshua L., Frank Barickman, John Martin, Sughosh Rao, Scott Schnelle, and Bowen Weng, “A Novel Method to Evaluate the Safety of Highly Automated Vehicles,” presentation at the 25th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration, Detroit, Mich., 2017.

FDA—See U.S. Food and Drug Administration.

Federal Aviation Administration, “System Design and Analysis,” Advisory Circular 25.1309-1A, June 21, 1988. As of July 20, 2020:

https://www.faa.gov/regulations_policies/advisory_circulars/index.cfm/go/document.information/documentid/22680

———, “Safety Management System (SMS): Components,” webpage, September 11, 2017. As of August 31, 2020:

<https://www.faa.gov/about/initiatives/sms/explained/components/>

———, “Safety Management System (SMS),” webpage, June 21, 2019. As of August 31, 2020: <https://www.faa.gov/about/initiatives/sms/>

Federal Highway Administration, *Surrogate Safety Assessment Model and Validation: Final Report*, McLean, Va.: U.S. Department of Transportation, February 2008. As of September 8, 2018:

<https://rosap.nhtl.bts.gov/view/dot/35896>

Federal Highway Administration, Policy and Governmental Affairs, Office of Highway Policy Information, “Highway Statistics Series: Highway Statistics 2017,” webpage, December 2018. As of August 30, 2020:

<https://www.fhwa.dot.gov/policyinformation/statistics/2017/d11c.cfm>

Fewtrell, Lorna, and Jamie Bartram, eds., *Water Quality: Guidelines, Standards and Health*, London: IWA, 2001. As of August 28, 2020:

<https://apps.who.int/iris/bitstream/handle/10665/42442/924154533X.pdf?sequence=1&isAllowed=y>

Finlay, Steve, “Uber Improves ‘Safety Culture’ in Aftermath of Fatal AV Accident,” WardsAuto, August 7, 2019. As of July 6, 2020:

<https://www.wardsauto.com/car-management-briefing-seminars/uber-improves-safety-culture-aftermath-fatal-av-accident>

Finucane, Melissa L., Ali Alhakami, Paul Slovic, and Stephen M. Johnson, “The Affect Heuristic in Judgments of Risks and Benefits,” *Journal of Behavioral Decision Making*, Vol. 13, No. 1, 2000, pp. 1–17.

Fischhoff, Baruch, Paul Slovic, Sarah Lichtenstein, Stephen Read, and Barbara Combs, “How Safe Is Safe Enough? A Psychometric Study of Attitudes Towards Technological Risks and Benefits,” *Policy Sciences*, Vol. 9, No. 2, 1978, pp. 127–152.

Fraade-Blonar, Laura, Marjory S. Blumenthal, James M. Anderson, and Nidhi Kalra, *Measuring Automated Vehicle Safety: Forging a Framework*, Santa Monica, Calif.: RAND Corporation,

- RR-2662, 2018. As of June 22, 2020:
https://www.rand.org/pubs/research_reports/RR2662.html
- Ghazizadeh, Mahtab, John D. Lee, and Linda Ng Boyle, “Extending the Technology Acceptance Model to Assess Automation,” *Cognition, Technology & Work*, Vol. 14, No. 1, 2012, pp. 39–49.
- Griffor, Edward R., Christopher Greer, and David A. Wollman, *Workshop Report: Consensus Safety Measurement for Automated Driving System-Equipped Vehicles*, Gaithersburg, Md.: National Institute of Standards and Technology, September 23, 2019. As of August 30, 2020:
<https://www.nist.gov/publications/workshop-report-consensus-safety-measurement-methodologies-automated-driving-system>
- Guldenmund, Frank W., “The Nature of Safety Culture: A Review of Theory and Research,” *Safety Science*, Vol. 34, 2000, pp. 215–257.
- Gupta, Nidhi, Arnout R. H. Fisher, and Lynn J. Frewer, “Socio-Psychological Determinants of Public Acceptance of Technologies: A Review,” *Public Understanding of Science*, March 1, 2011, pp. 782–795. As of June 28, 2020 available at:
<https://journals.sagepub.com/doi/full/10.1177/0963662510392485>
- Hahn, Susan E., and Lawrence R. Murphy, “A Short Scale for Measuring Safety Climate,” *Safety Science*, Vol. 46, 2008, pp. 1047–1066.
- Hartwich, Franziska, Claudia Witzlack, Matthias Beggiato, and Josef F. Krems, “The First Impression Counts—A Combined Driving Simulator and Test Track Study on the Development of Trust and Acceptance of Highly Automated Driving,” *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 65, August 2019, pp. 522–535.
- Hayward, John C., “Near Miss Determination Through Use of a Scale of Danger,” in *Proceedings of the 51st Annual Meeting of the Highway Research Board*, Washington, D.C., 1972, pp. 24–35.
- Hollnagel, Erik, Robert L. Wears, and Jeffrey Braithwaite, *From Safety-I to Safety-II*, University of Southern Denmark, University of Florida, and Macquarie University, white paper, 2015. As of June 28, 2020:
<https://www.england.nhs.uk/signuptosafety/wp-content/uploads/sites/16/2015/10/safety-1-safety-2-white-papr.pdf>
- IEEE—See Institute of Electrical and Electronics Engineers.
- Institute of Electrical and Electronics Engineers, “WG: VT/VTs/AV Decision Making,” webpage, undated. As of August 31, 2020:
<https://sagroups.ieee.org/2846/>

- Institute of Electrical and Electronics Engineers Standards Association, “P2040—Standard for General Requirements for Fully Automated Vehicles Driving on Public Roads,” webpage, undated-a. As of August 31, 2020:
<https://standards.ieee.org/project/2040.html>
- , “P2846—A Formal Model for Safety Considerations in Automated Vehicle Decision Making,” webpage, undated-b. As of August 30, 2020:
<https://standards.ieee.org/project/2846.html>
- Insurance Institute for Highway Safety and Highway Loss Data Institute, homepage, undated. As of August 31, 2020:
<https://www.iihs.org/>
- , “Red Light Running,” webpage, February 2020. As of August 31, 2020:
<https://www.iihs.org/topics/red-light-running>
- International Organization for Standardization, “Road Vehicles—Safety and Cybersecurity for Automated Driving Systems—Design, Verification and Validation Methods,” ISO/CD TR 4804, undated. As of August 30, 2020:
<https://www.iso.org/standard/80363.html>
- , “Road Vehicles—Functional Safety—Part 1: Vocabulary,” ISO 26262-1:2011, 2011. As of August 30, 2020:
<https://www.iso.org/standard/43464.html>
- , “Road Vehicles—Safety of the Intended Functionality,” ISO/PAS 21448:2019, 2019. As of August 30, 2020:
<https://www.iso.org/standard/70939.html>
- ISO—See International Organization for Standardization.
- Jiang, Lixin, Lindsey M. Lavaysse, and Tahira M. Probst, “Safety Climate and Safety Outcomes: A Meta-Analytic Comparison of Universal vs. Industry-Specific Safety Climate Predictive Validity,” *Work & Stress*, Vol. 33, No. 1, 2019, pp. 41–57.
- Johnson, Judith A., *FDA Regulation of Medical Devices*, Washington, D.C.: Congressional Research Service, R42130, September 2016. As of August 30, 2020:
<https://fas.org/sgp/crs/misc/R42130.pdf>
- Johnsson, Carl, Aliaksei Laureshyn, and Tim De Ceunynck, “In Search of Surrogate Safety Indicators for Vulnerable Road Users: A Review of Surrogate Safety Indicators,” *Transportation Reviews*, Vol. 38, No. 5, 2018.
- Junietz, Philipp Matthias, *Microscopic and Macroscopic Risk Metrics for the Safety Validation of Automated Driving*, doctoral thesis, Darmstadt: Technische Universität, 2019. As of June

28, 2020:

<https://tuprints.ulb.tu-darmstadt.de/9282/>

Junietz, Philipp, Udo Steininger, and Hermann Winner, “Macroscopic Safety Requirements for Highly Automated Driving,” *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2673, No. 3, 2019.

Kahneman, Daniel, *Thinking, Fast and Slow*, New York: Macmillan, 2011.

Kalra, Nidhi, and David G. Groves, *The Enemy of Good: Estimating the Cost of Waiting for Nearly Perfect Automated Vehicles*, Santa Monica, Calif.: RAND Corporation, RR-2150-RC, 2017. As of August 30, 2020:

https://www.rand.org/pubs/research_reports/RR2150.html

Kalra, Nidhi, and Susan M. Paddock, *Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?* Santa Monica, Calif.: RAND Corporation, RR-1478-RC, 2016. As of June 29, 2020:

https://www.rand.org/pubs/research_reports/RR1478.html

Kines, Pete, Jorma Lappalainen, Kim Lyngby Mikkelsen, Espen Olsen, Anders Pousette, Jorunn Tharaldsen, Kristinn Tómasson, and Marianne Törnerd, “Nordic Safety Climate Questionnaire (NOSACQ-50): A New Tool for Diagnosing Occupational Safety Climate,” *International Journal of Industrial Ergonomics*, Vol. 41, No. 6, November 2011, pp. 634–646.

Klauer, Sheila G., Feng Guo, Bruce G. Simons-Morton, Marie Claude Ouimet, Susan E. Lee, and Thomas A. Dingus, “Distracted Driving and Risk of Road Crashes Among Novice and Experienced Drivers,” *New England Journal of Medicine*, Vol. 370, No. 1, 2014, pp. 54–59.

Koopman, Philip, Beth Osyk, and Jack Weast, *Autonomous Vehicles Meet the Physical World: RSS, Variability, Uncertainty, and Proving Safety*, arxiv.org, 2019. As of June 28, 2020: <https://arxiv.org/ftp/arxiv/papers/1911/1911.01207.pdf>

Koopman, Philip, and Michael Wagner, *Toward a Framework for Highly Automated Vehicle Validation*, SAE International, Technical Paper 2018-01-1071, April 3, 2018.

Kornhauser, Alain, “8.25-Birthday-0612420,” *Smart Driving Cars*, newsletter, June 12, 2020. As of August 30, 2020:

<https://smartdrivingcar.com/8-25-birthday-0612420/>

Lambregts, Antonius A., “Flight Envelope Protection for Automatic and Augmented Manual Control,” in *Proceedings of the EutoGNC 2013, 2nd CEAS Specialist Conference on Guidance, Navigation and Control*, 2013, pp. 1364–1383.

- Liu, Peng, Yong Du, and Zhigang Xu, “Machines Versus Humans: People’s Biased Responses to Traffic Accidents Involving Self-Driving Vehicles,” *Accident Analysis & Prevention*, Vol. 125, 2019, pp. 232–240.
- Liu, Peng, Zhang Yawen, and He Zhen, “The Effect of Population Age on the Acceptability of Safety of Self-Driving Vehicles,” *Reliability Engineering & System Safety*, Vol. 185, Iss. C, 2019, pp. 341–347.
- Lund, Frederick Hansen, “The Psychology of Belief: A Study of Its Emotional and Volitional Determinants,” *Journal of Abnormal and Social Psychology*, Vol. 20, No. 1, 1925, pp. 63–81.
- Lyft, “Level 5 Open Data: Advancing Self-Driving Technology, Together,” data set, undated. As of June 24, 2020:
<https://level5.lyft.com/dataset/>
- Mahmud, S. M. Sohel, Luis Ferreira, Shamsul Hoque, and Ahmad Tavassoli, “Application of Proximal Surrogate Indicators for Safety Evaluation: A Review of Recent Developments and Research Needs,” *IATSS Research*, Vol. 41, No. 4, December 2017, pp. 153–163. As of September 7, 2018:
<https://www.sciencedirect.com/science/article/pii/S0386111217300286>
- Manfreda, Anton, Klara Ljubi, and Aleš Groznik, “Autonomous Vehicles in the Smart City Era: An Empirical Study of Adoption Factors Important for Millennials,” *International Journal of Information Management*, December 2019.
- Marshall, Aarian, “What Can the Trolley Problem Teach Self-Driving Car Engineers?” *Wired*, October 24, 2018. As of August 28, 2020:
<https://www.wired.com/story/trolley-problem-teach-self-driving-car-engineers/>
- McDonald, Ashley, Cher Carney, and Daniel V. McGehee, *Vehicle Owners’ Experiences with and Reactions to Advanced Driver Assistance Systems*, Washington, D.C.: AAA Foundation for Traffic Safety, September 2018. As of September 4, 2020:
https://aaaafoundation.org/wp-content/uploads/2018/09/VehicleOwnersExperiencesWithADAS_TechnicalReport.pdf
- Melchers, Robert E., “On the ALARP Approach to Risk Management,” *Reliability Engineering & System Safety*, Vol. 71, No. 2, February 2001, pp. 201–208.
- Menon, Nikhil, Yu Zhang, Abdul Rawoof Pinjari, and Fred Mannering, “A Statistical Analysis of Consumer Perceptions Towards Automated Vehicles and Their Intended Adoption,” *Transportation Planning and Technology*, Vol. 43, No. 3, 2020, pp. 253–278.
- Michaels, Dave, Andy Pasztor, and Andrew Tangel, “Two FAA Officials Are Key Witnesses in Criminal Probe of Ex-Boeing Pilot,” *Wall Street Journal*, March 14, 2020.

- Mobileye, “Responsibility-Sensitive Safety,” webpage, undated. As of August 25, 2020:
<https://www.mobileye.com/responsibility-sensitive-safety/>
- Möller, Niklas, Sven Ove Hansson, and Martin Peterson, “Safety Is More Than the Antonym of Risk,” *Journal of Applied Philosophy*, Vol. 23, No. 4, 2006, pp. 419–432.
- Moody, Joanna, Nathaniel Bailey, and Jinhua Zhao, “Public Perceptions of Autonomous Vehicle Safety: An International Comparison,” *Safety Science*, Vol. 121, 2020, pp. 634–650.
- Nadvi, Khalid, “Global Standards, Global Governance, and the Organization of Global Value Chains,” *Journal of Economic Geography*, Vol. 8, 2008, pp. 323–343.
- National Governors Association Center for Best Practices, *State Public Safety and Autonomous Vehicle Technology*, Washington, D.C.: National Governors Association, 2018. As of August 30, 2020:
<https://www.nga.org/wp-content/uploads/2019/09/Autonomous-Vehicle-Technology.pdf>
- National Highway Traffic Safety Administration, “New Test Tracking Tool,” webpage, undated-a. As of August 20, 2020:
<https://www.nhtsa.gov/automated-vehicles-safety/av-test>
- , “Ratings,” webpage, undated-b. As of June 23, 2020:
<https://www.nhtsa.gov/ratings>
- , “Regulations,” webpage, undated-c. As of August 30, 2020:
<https://www.nhtsa.gov/laws-regulations/fmvss>
- , “Voluntary Safety Self-Assessment,” webpage, undated-d. As of August 30, 2020:
<https://www.nhtsa.gov/automated-driving-systems/voluntary-safety-self-assessment>
- , *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*, Washington, D.C.: U.S. Department of Transportation, DOT HS 810 594, 2006. As of August 29, 2020:
<https://vtechworks.lib.vt.edu/bitstream/handle/10919/55090/DriverInattention.pdf>
- , “Police-Reported Motor Vehicle Traffic Crashes in 2017,” Traffic Safety Facts Research Note, July 2019a. As of August 31, 2020:
<https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812696>
- , “NHTSA Announces Coming Upgrades to New Car Assessment Program,” news release, October 16, 2019b. As of September 9, 2020:
<https://www.nhtsa.gov/press-releases/ncap-upgrades-coming>
- , “Nuro, Inc.: Grant of Temporary Exemption for a Low-Speed Vehicle with an Automated Driving System,” U.S. Department of Transportation, Docket No. NHTSA-2019-0017, *Federal Register*, Vol. 85, No. 7826, February 11, 2020a. As of June 23, 2020:

- <https://www.federalregister.gov/documents/2020/02/11/2020-02668/nuro-inc-grant-of-temporary-exemption-for-a-low-speed-vehicle-with-an-automated-driving-system>
- , “Agency Information Collection Activities; Notice and Request for Comment; Government 5-Star Safety Ratings Label Consumer Research,” *Federal Register*, Vol. 85, No. 23598, April 28, 2020b, pp. 23598-23600. As of September 9, 2020:
<https://www.federalregister.gov/documents/2020/04/28/2020-08949/agency-information-collection-activities-notice-and-request-for-comment-government-5-star-safety>
- , “Early Estimates of 2019 Motor Vehicle Traffic Data Show Reduced Fatalities for Third Consecutive Year,” press release, May 5, 2020c. As of August 30, 2020:
<https://www.nhtsa.gov/press-releases/early-estimates-traffic-fatalities-2019>
- , “Traffic Safety Facts Annual Report Tables,” webpage, June 30, 2020d. As of August 31, 2020:
<https://cdan.nhtsa.gov/tsftables/tsfar.htm#>
- National Household Travel Survey, *Developing a Best Estimate of Annual Vehicle Mileage for 2017 NHTS Vehicles*, 2017. As of August 31, 2020:
https://nhts.ornl.gov/assets/2017BESTMILE_Documentation.pdf
- National Institute of Standards and Technology, “Cyber-Physical Systems,” webpage, undated. As of August 30, 2020:
<https://www.nist.gov/el/cyber-physical-systems>
- National Research Council, *Software for Dependable Systems: Sufficient Evidence?* Washington, D.C.: National Academies Press, 2007.
- National Safety Council, “Happy Anniversary, Road to Zero!” *Safety First*, blog post, October 23, 2017. As of August 20, 2020:
<https://www.nsc.org/safety-first-blog/happy-anniversary-road-to-zero-1>
- National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrians*, Tempe, Ariz., March 18, 2018, Washington, D.C., November 19, 2019a. As of August 30, 2020:
<https://www.nts.gov/investigations/AccidentReports/Reports/HAR1903.pdf>
- , “‘Inadequate Safety Culture’ Contributed to Uber Automated Test Vehicle Crash—NTSB Calls for Federal Review Process for Automated Vehicle Testing on Public Roads,” news release, November 19, 2019b. As of August 30, 2020:
<https://www.nts.gov/news/press-releases/Pages/NR20191119c.aspx>
- Nelkin, Dorothy, “Communicating Technological Risk: The Social Construction of Risk Perception,” *Annual Review of Public Health*, Vol. 10, No. 1, 1989, pp. 95–113.
- NHTSA—See National Highway Traffic Safety Administration.

- Nordgren, Loran F., Joop van der Plight, and Frenk van Harreveld, “Unpacking Perceived Control in Risk Perception: The Mediating Role of Anticipated Regret,” *Journal of Behavioral Decision Making*, Vol. 20, No. 5, 2007, pp. 533–544.
- NSC—See National Safety Council.
- NTSB—See National Transportation Safety Board.
- NVIDIA, “Planning a Safer Path,” webpage, undated. As of August 20, 2020:
<https://www.nvidia.com/en-us/self-driving-cars/safety-force-field/>
- Oehler, Andreas, and Stefan Wendt, “Good Consumer Information: The Information Paradigm at Its (Dead) End?” *Journal of Consumer Policy*, Vol. 40, No. 2, 2017, pp. 179–191.
- Oltedal, Sigve, Bjorn-Elin Moen, Hroar Klempe, and Torbjørn Rundmo, “Explaining Risk Perception: An Evaluation of Cultural Theory,” *Trondheim: Norwegian University of Science and Technology*, Vol. 85, 2004.
- Osephchuck, John M., “The History of the Microwave Oven: A Critical Review,” *2009 IEEE MMT-S International Microwave Symposium Digest*, 2009, pp. 1397–1400.
- Partners for Automated Vehicle Education, “PAVE Poll: Americans Wary of AVs but Say Education and Experience with Technology Can Build Trust,” webpage, undated. As of June 19, 2020:
<https://pavecampaing.org/news/pave-poll-americans-wary-of-avs-but-say-education-and-experience-with-technology-can-build-trust/>
- , “PAVE Poll: Fact Sheet,” May 2020. As of June 19, 2020:
https://pavecampaing.org/wp-content/uploads/2020/05/PAVE-Poll_Fact-Sheet.pdf
- Pegasus, “Requirements & Conditions—Stand 7: Social Acceptance for HAD (L3),” presentation at Pegasus Symposium 2019, University of Glasgow, May 14, 2019. As of June 20, 2020:
https://www.pegasusprojekt.de/files/tmpl/PDF-Symposium/07_Social-Acceptance-for-HAD.pdf
- Pollard, Michael, and Matthew D. Baird, *The RAND American Life Panel: Technical Description*, Santa Monica, Calif.: RAND Corporation, RR-1651, 2017. As of September 20, 2020:
https://www.rand.org/pubs/research_reports/RR1651.html
- Rae, Andrew J., *Acceptable Residual Risk—Principles, Philosophies and Practicalities*, paper presented at second Institution of Engineering and Technology International Conference on System Safety, London, 2007.

- Rahman, Md Mahmudur, Mary F. Lesch, William J. Horrey, and Lesley Strawderman, "Assessing the Utility of TAM, TPB, and UTAUT for Advanced Driver Assistance Systems," *Accident Analysis & Prevention*, Vol. 108, 2017, pp. 361–373.
- RAND Corporation, American Life Panel, "Welcome to the ALP Data Pages," webpage, undated. As of September 9, 2020:
<https://alpdata.rand.org>
- Robinson, Lisa A., and James K. Hammitt, "Behavioral Economics and Regulatory Analysis," *Risk Analysis: An International Journal*, Vol. 31, No. 9, 2011, pp. 1408–1422.
- Rovira, Ericka, Anne Collins McLaughlin, Richard Pak, and Luke High, "Looking for Age Differences in Self-Driving Vehicles: Examining the Effects of Automation Reliability, Driving Risk, and Physical Impairment on Trust," *Frontiers in Psychology*, Vol. 10, 2019, pp. 1–13.
- Rundmo, Torbjørn, Trond Nordfjærn, Hilde Hestad Iversen, Sigve Olteidal, and Stig H. Jørgensen, "The Role of Risk Perception and Other Risk-Related Judgements in Transportation Mode Use," *Safety Science*, Vol. 49, No. 2, 2011, pp. 226–235.
- SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," Standard J3016_201806, June 15, 2018. As of June 28, 2020:
https://www.sae.org/standards/content/j3016_201806/
- Santacreu, Alexandre, George Yannis, Omblin de Saint Léon, and Philippe Crist, *Safe Micromobility*, Paris, France: International Transport Forum, 2020.
- Sax, Joanna K., and Neal Doran, "Ambiguity and Consumer Perceptions of Risk in Various Areas of Biotechnology," *Journal of Consumer Policy*, Vol. 42, No. 1, 2019, pp. 47–58.
- Schäbe, Hendrik, *Different Principles Used for Determination of Tolerable Hazard Rates*, Cologne, Germany: Institute for Software, Electronics, Railroad Technology, 2001. As of August 20, 2020:
<http://www.railway-research.org/IMG/pdf/041.pdf>
- Scholderer, Joachim, and Lynn J. Frewer, "The Biotechnology Communication Paradox: Experimental Evidence and the Need for a New Strategy," *Journal of Consumer Policy*, Vol. 26, 2003, pp. 125–157.
- Schulze, William D., "Ethics, Economics and the Value of Safety," in *Societal Risk Assessment*, Boston, Mass.: Springer, 1980, pp. 217–231.
- Sener, Ipek N., Johanna Zmud, and Chris Simek, *Examining Future Automated Vehicle Usage: A Focus on the Role of Ride Hailing*, College Station, Tex.: Texas A&M Transportation Institute, May 2018.

- Sener, Ipek N., Johanna Zmud, and Thomas Williams, “Measures of Baseline Intent to Use Automated Vehicles: A Case Study of Texas Cities,” *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 62, April 2019, pp. 66–77.
- Shalev-Schwartz, Shai, Shaked Shammah, and Amnon Shashua, *On a Formal Model of Safe and Scalable Self-Driving Cars*, arxiv.org, 2017. As of June 28, 2020: <https://arxiv.org/pdf/1708.06374.pdf>
- Silberg, Gary, *Islands of Autonomy*, Amstelveen, Netherlands: KPMG, 2017. As of June 25, 2020: <https://assets.kpmg/content/dam/kpmg/za/pdf/2017/11/islands-of-autonomy-web.pdf>
- Slovic, Paul, “Perceptions of Risk: Reflections on the Psychometric Paradigm,” in Sheldon Krinsky and Dominic Golding, eds., *Social Theories of Risk*, New York: Praeger, 1990, pp. 117–152.
- Slovic, Paul, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor, “Risk as Analysis and Risk as Feelings: Some Thoughts About Affect, Reason, Risk, and Rationality,” *Risk Analysis*, Vol. 24, No. 2, 2004, pp. 311–322.
- Smith, Bryant Walker, “How Reporters Can Evaluate Automated Driving Announcements,” *Journal of Law and Mobility*, April 19, 2020. As of August 30, 2020: <https://futurist.law.umich.edu/author/bryantws/>
- Souders, Dustin J., Ryan Best, and Neil Charness, “Valuation of Active Blind Spot Detection Systems by Younger and Older Adults,” *Accident Analysis & Prevention*, Vol. 106, 2017, pp. 505–514.
- Stamatiadis, Nick, and Arnold Stromberg, “Crash Involvement of Drivers with Multiple Crashes,” *Accident Analysis & Prevention*, Vol. 38, No. 3, June 2006.
- Stapelberg, Rudolph Frederick, “Safety and Risk in Engineering Design,” *Handbook of Reliability, Availability, Maintainability and Safety in Engineering Design*, London: Springer, 2009, pp. 529–798.
- Starr, Chauncey, “Social Benefit Versus Technological Risk,” *Science*, Vol. 165, No. 3899, 1969, pp. 1232–1238.
- Steffl-Mabry, Joette, “A Social Judgment Analysis of Information Source Preference Profiles: An Exploratory Study to Empirically Represent Media Selection Patterns,” *Journal of the American Society for Information Science and Technology*, Vol. 54, No. 9, 2003, pp. 879–904.
- Stewart, Emily, “Self-Driving Cars Have to Be Safer Than Regular Cars. The Question Is How Much,” *Vox*, May 17, 2019. As of June 28, 2020:

<https://www.vox.com/recode/2019/5/17/18564501/self-driving-car-morals-safety-tesla-waymo>

Stipancic, Joshua, Luis Miranda-Moreno, and Nicolas Saunier, “Vehicle Manoeuvres as Surrogate Safety Measures: Extracting Data from the GPS-Enabled Smartphones of Regular Drivers,” *Accident Analysis & Prevention*, Vol. 115, June 2018, pp. 160–169.

Stone, Vernon A., “A Primacy Effect in Decision-Making by Jurors,” *Journal of Communication*, Vol. 19, No. 3, 1969, pp. 239–247.

Taleb, Nassim Nicholas, *Antifragile: Things That Gain from Disorder*, New York: Random House, 2014.

Teoh, Eric R., “What’s in a Name? Drivers’ Perceptions of the Use of Five SAE Level 2 Driving Automation Systems,” *Journal of Safety Research*, Vol. 72, 2020, pp. 145–151.

Tesla, “Future of Driving,” webpage, undated. As of August 30, 2020:
<https://www.tesla.com/autopilot>

Thong, James Y. L., and Xin Xu, “Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology,” *MIS Quarterly*, Vol. 36, 2012, pp. 157–178.

Tversky, Amos, and Daniel Kahneman, “The Framing of Decisions and the Psychology of Choice,” *Science*, Vol. 211, No. 4481, 1981, pp. 453–458.

Uber Advanced Technologies Group, “Safety Case Framework,” website, undated. As of August 30, 2020:
<https://www.uber.com/us/en/atg/safety/safety-case-framework/>

———, *A Principled Approach to Safety: Safety Report 2020*, August 2020. As of September 28, 2020:
<https://uber.app.box.com/v/UberATGSafetyReport>

Uber ATG Safety Team, “Uber ATG Releases 2020 Safety Report,” *Medium*, August 28, 2020. As of September 28, 2020:
<https://medium.com/@UberATG/uber-atg-releases-2020-safety-report-575db33f2bd7>

Uchida, Nobuyuki, Maki Kawakoshi, Takashi Tagawa, and Tsutomu Mochida, “An Investigation of Factors Contributing to Major Crash Types in Japan Based on Naturalistic Driving Data,” *IATSS Research*, Vol. 34, No. 1, 2010, pp. 22–30.

Underwriters Laboratories, “About Us,” webpage, undated-a. As of August 20, 2020:
<https://ul.org/about>

- , “Presenting the Standard for Safety for the Evaluation of Autonomous Vehicles and Other Products,” webpage, undated-b. As of July 20, 2020:
<https://ul.org/UL4600>
- Underwriters Laboratories Standards, “Standard for Evaluation of Autonomous Products,” Standard 4600, Edition 1, April 1, 2020.
- Urmson, Chris, “Putting Safety into Practice: Aurora’s Safety Approach,” Aurora blog post, October 2, 2019. As of July 20, 2020:
<https://medium.com/aurora-blog/putting-safety-into-practice-auroras-safety-approach-5297de2d8276>
- U.S. Census Bureau and U.S. Bureau of Labor Statistics, “Current Population Survey (CPS),” webpage, undated. As of July 20, 2020:
<https://www.census.gov/programs-surveys/cps.html>
- U.S. Food and Drug Administration, “Medical Devices,” webpage, undated. As of August 30, 2020:
<https://www.fda.gov/Medical-Devices>
- , “Digital Health Software Precertification (Pre-Cert) Program,” webpage, July 18, 2019a. As of August 30, 2020:
<https://www.fda.gov/medical-devices/digital-health/digital-health-software-precertification-pre-cert-program>
- , “Generally Recognized as Safe (GRAS),” webpage, September 6, 2019b. As of August 28, 2020:
<https://www.fda.gov/food/food-ingredients-packaging/generally-recognized-safe-gras>
- , “Digital Health,” webpage, August 27, 2020. As of August 30, 2020:
<https://www.fda.gov/medical-devices/digital-health>
- U.S. House of Representatives, Committee on Transportation and Infrastructure, *The Boeing 737 MAX Aircraft: Costs, Consequences, and Lessons from Its Design, Development, and Certification*, final committee report, September 2020. As of October 2, 2020:
<https://transportation.house.gov/imo/media/doc/2020.09.15%20FINAL%20737%20MAX%20Report%20for%20Public%20Release.pdf>
- Venkatesh, Viswanath, Michael G. Morris, Gordon B. Davis, and Fred D. Davis, “User Acceptance of Information Technology: Toward a Unified View,” *MIS Quarterly*, Vol. 27, No. 3, 2003, pp. 425–478.
- Venkatesh, Viswanath, James Y. L. Thong, and Xin Xu, “Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology,” *MIS Quarterly*, Vol. 36, No. 1, 2012, pp. 157–178.

- Vermont Agency of Transportation, Policy, Planning, and Intermodal Development Division, *Vermont Automated Vehicle Testing Permit: Guidance and Application*, Barre, Vt.: Vermont Agency of Transportation, April 24, 2020. As of August 30, 2020:
<https://vtrans.vermont.gov/sites/aot/files/planning/documents/VT%20AV%20Testing%20Guidance%20and%20Application.042420.pdf>
- Vierendeels, Geert, Genserik Reniers, Karolien van Nunen, and Koen Ponnet, “An Integrative Conceptual Framework for Safety Culture: The Egg Aggregated Model (TEAM) of Safety Culture,” *Safety Science*, Vol. 103, 2018, pp. 323–339.
- Virginia Tech Transportation Institute, *InSight Data Access*, website, undated (registration required). As of August 20, 2020:
<https://insight.shrp2nds.us/>
- Viscusi, W. Kip, *Regulation of Health, Safety, and Environmental Risks*, Cambridge, Mass.: National Bureau of Economic Research, Working Paper 11934, 2006. As of June 16, 2020:
<https://www.nber.org/papers/w11934>
- Ward, Carley, Martina Raue, Chaiwoo Lee, Lisa D-Ambrosio, and Joseph F. Coughlin, “Acceptance of Automated Driving Across Generations: The Role of Risk and Benefit Perception, Knowledge, and Trust,” in Masaaki Kurosu, ed., *Human-Computer Interaction: User Interface Design, Development, and Multimodality 19th International Conference, HCI International 2017 Proceeding, Part I*, New York: Springer International Publishing, 2017, pp. 254–266.
- Waymo, “Waymo Open Dataset,” data set, undated. As of June 24, 2020:
<https://waymo.com/open/>
- Weast, Jack, “Metrics and Assumptions in Safety Assurance,” presentation, July 29, 2020. As of July 20, 2020:
<http://sagroups.ieee.org/2846/wp-content/uploads/sites/124/2020/08/Metrics-and-Assumptions-in-Safety-Assurance-7-29-20-1.pdf>
- Weinstein, Neil D., “Why It Won’t Happen to Me: Perceptions of Risk Factors and Susceptibility,” *Health Psychology*, Vol. 3, No. 5, 1984, pp. 431–457.
- Weng, Bowen, Sughosh Rao, Eeshan Deosthale, Scott Schnelle, and Frank Barickman, *Model Predictive Instantaneous Safety Metric for Evaluation of Automated Driving Systems*, arxiv.org, May 2020. As of June 28, 2020:
<https://arxiv.org/pdf/2005.09999.pdf>
- Winkelman, Zev, Maya Buenaventura, James M. Anderson, Nahom M. Beyene, Pavan Katkar, and Gregory Cyril Baumann, *When Autonomous Vehicles Are Hacked, Who Is Liable?* Santa Monica, Calif.: RAND Corporation, RR-2654-RC, 2019. As of September 4, 2020:
https://www.rand.org/pubs/research_reports/RR2654.html

- Wishart, Jeffrey, Steven Como, Maria Elli, Brendan Russo, Jack Weast, Niraj Altekar, and Emmanuel James, *Driving Safety Performance Assessments Metrics for ADS-Equipped Vehicles*, SAE International, Technical Paper 2020-01-1206, 2020.
- Wohl, Jennifer B., “Consumer’s Decision-Making and Risk Perceptions Regarding Foods Produced with Biotechnology,” *Journal of Consumer Policy*, Vol. 21, 1998, pp. 387–404.
- Wu, Kun-Feng, Jonathan Aguero-Valverde, and Paul P. Jovanis, “Using Naturalistic Driving Data to Explore the Association Between Traffic Safety–Related Events and Crash Risk at Driver Level,” *Accident Analysis & Prevention*, Vol. 72, November 2014, pp. 210–218.
- Yoshida, Junko, “Can Mobileye Validate ‘True Redundancy’?” *EE Times*, May 22, 2018. As of June 28, 2020:
<https://www.eetimes.com/can-mobileye-validate-true-redundancy/>
- , “AV Safety Ventures Beyond ISO 26262,” *EE Times*, March 5, 2019. As of June 28, 2020:
<https://www.eetimes.com/av-safety-ventures-beyond-iso-26262/#>
- , “Safe Autonomy: UL 4600 and How It Grew,” *EE Times*, April 2, 2020. As of June 28, 2020 available at:
<https://www.eetimes.com/safe-autonomy-ul-4600-and-how-it-grew/>
- Zendrive, “Mobility Amidst Lockdown: Every Minute on the Road Is Riskier,” webpage, May 2020. As of August 31, 2020:
<https://live.zendrive.com/covid-study>
- Zhang, Qiaoning, X. Jessie Yang, and Lionel Peter Robert, “Expectations and Trust in Automated Vehicles,” *CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, April 2020, pp. 1–9.



Establishing whether automated vehicles (AVs) are acceptably safe is not straightforward, and continual technology modification adds complication. RAND Corporation researchers analyzed three categories of approach—measurements, processes, and thresholds—and noted the different kinds of evidence associated with each, the ways in which different approaches can be used together, and the degree to which stakeholder groups agree on the merits of these approaches. This report complements discussion of measurement and analytical issues with a discussion of challenges in communicating about AV safety, especially to the general public. Its recommendations are aimed at both industry and government.

\$39.00

ISBN-10 1-9774-0603-3
ISBN-13 978-1-9774-0603-3



www.rand.org