From driverless dilemmas to more practical commonsense tests for automated vehicles

Julian De Freitas[a,1], Andrea Censi[b], Bryant Walker Smith[c,d], Luigi Di Lillo[e], Sam E. Anthony[f], and Emilio Frazzoli[b]

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved January 17, 2021 (received for review June 5, 2020)

For the first time in history, automated vehicles (AVs) are being deployed in populated environments. This unprecedented transformation of our everyday lives demands a significant undertaking: endowing complex autonomous systems with ethically acceptable behavior. We outline how one prominent, ethically relevant component of AVs—driving behavior—is inextricably linked to stakeholders in the technical, regulatory, and social spheres of the field. Whereas humans are presumed (rightly or wrongly) to have the "common sense" to behave ethically in new driving situations beyond a standard driving test, AVs do not (and probably should not) enjoy this presumption. We examine, at a high level, how to test the common sense of an AV. We start by reviewing discussions of "driverless dilemmas," adaptions of the traditional "trolley dilemmas" of philosophy that have sparked discussion on AV ethics but have limited use to the technical and legal spheres. Then, we explain how to substantially change the premises and features of these dilemmas (while preserving their behavioral diagnostic spirit) in order to lay the foundations for a more practical and relevant framework that tests driving common sense as an integral part of road rules testing.

automated driving | public health | ethics | policy | artificial intelligence

Ethics Is an Integral Challenge for the Introduction of Automated Vehicles

Each year around the world, motor vehicle crashes kill 1.25 million people and injure another 20 million. Human errors—and the systems that make it so easy for these errors to occur and be dangerous (1, 2)—are a cause of at least 90% of these crashes (3). If new technology could prevent these deaths, it would improve public health at a scale on par with vaccines or penicillin.

One of the main objectives of automated driving technologies is to reduce these casualties, by creating safer traffic situations rather than only reacting to dangerous ones as they occur. However, the societally successful introduction of automated vehicles (AVs) demands the resolution of technical, regulatory, and social challenges (4). Ethics is at the intersection of these challenges: The software of AVs will produce behaviors with ethical ramifications (technical); governments will often seek to align their laws with norms of ethical behavior (legal); and the public's perception of the ethics of automation may affect their acceptance of the technologies (social). How can we ethically address these challenges in these contexts?

In this paper, we take a pragmatic approach. First, we outline how ethics in a broader sense is integral to various stakeholders in the technical, regulatory, and social spheres of the AV industry (*Ethics Is Ubiquitous in Automated Driving*). Then, we consider the idea of "driverless dilemmas," which have sparked recent discussion on the ethics of AVs (*A Narrow View of This Ethical Landscape: Driverless Dilemmas*). Finally, we discuss how these dilemmas can be modified to inspire a more practical, ethical testing framework for AVs and why this is needed (*A Pragmatic Implication: Test for Commonsense Behavior*), before providing recommendations for how to think of testing as more of a process than a singular event (*The Road Ahead:*

[a]Department of Psychology, Harvard University, Cambridge, MA 02138; [b]Department of Mechanical and Process Engineering, ETH Zürich, 8092 Zurich, Switzerland; [c]School of Law, University of South Carolina, Columbia, SC 29201; [d]College of Engineering and Computing, University of South Carolina, Columbia, SC 29201; [e]Property and Casualty Solutions, Reinsurance, Swiss Reinsurance Company, Ltd., Zurich 8022, Switzerland; and [f]Perceptive Automata Inc., Boston, MA 02108

Author contributions: J.D.F., A.C., B.W.S., L.D.L., S.E.A., and E.F. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: defreitas@g.harvard.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2010202118/-/DCSupplemental.

Published March 1, 2021.

Ethics as a Marriage, Not a Wedding). Throughout, we focus on the ethical component of dynamic driving behavior (such as speed and maneuvering). In particular, how should developers, regulators, and others ensure that AVs exhibit ethically relevant "common sense" by behaving in a way that is safe, predictable, reasonable, uniform, comfortable, and explainable under real-world conditions and constraints? We do not discuss broader ethical issues related to strategic aspects of driving and to transport more generally (such as whether, where, and how to travel) (4–6).

Ethics Is Ubiquitous in Automated Driving

We take for granted that drivers must take reasonable precautions to prevent unreasonable risks of driving. A risk is the product of a harm's probability and its severity, where the total risk of any particular course of action or inaction is the sum of all estimated risks associated with it (7, 8). Preventing unreasonable risk is optimal, and ethics is an approach to deciding how to act optimally by appealing to a set of guidelines that the applied sciences themselves cannot provide—such as values, principles, beliefs, norms, and purposes (4). People appeal to ethics in trying to categorically adhere to a principle like "never discriminate between individuals based on their social class" or when unconsciously (and probably imperfectly) weighing various risks to decide whether to speed to work (9, 10). In the case of AVs, ethical considerations serve as inputs to decisions that determine their behavior. For instance, whether a manufacturer prioritizes reducing harm over reducing liability depends in part on its internal commitment to safety as well as on its susceptibility to external regulatory, market, and reputational forces (11). In this way, the behavioral ethics challenges of AVs may seem like a Pandora's box, since they often require arbitrating among various stakeholders across the technical, legal, and social spheres. No single actor knows enough to solve all the issues on its own, any actor may miss the ethical relevance of a problem, and the solution devised by one actor may run counter to the interests of another. For example, regulators may not fully understand the underlying engineering challenges, and engineers may not understand the legal implications of implementing a given solution to a technical challenge.

In *SI Appendix*, we indicate how ethics is relevant to different stakeholders in the technical, legal, and social spheres. Each of these spheres (including marketing, which we do not discuss here) may also vary across space and evolve over time. For example, norms about compliance with traffic laws, dynamics between motorists and vulnerable road users, and conventional driving courtesies may all produce different conditions and expectations for AVs. Given these complexities, it is tempting to refrain from thinking about the ethical aspects of AV behavior altogether. However, when the AVs are on our roads, they will do something, and what and how they do it will have ethical relevance. Whereas humans can only be incentivized to drive a certain way, AVs can be made to do so.

A Narrow View of This Ethical Landscape: Driverless Dilemmas

One approach to start a discussion about the behavioral ethics of AVs are so-called driverless dilemmas, an idea inspired from "trolley dilemmas" in philosophy. The original trolley dilemma (12) asks you to imagine that a trolley is on course to hit five unsuspecting workers on the track, unless you redirect it to another track with only one worker on it. The dilemmas were originally designed to contrast different moral philosophies (13–15) and have also been imported by psychologists to study how people ordinarily make moral judgments (14, 16).

Recently, trolley dilemmas have been extended into the domain of AVs (17, 18), to suggest that AVs will face driverless dilemmas. A prominent subset of this work asks people on the web to consider simple scenarios in which a hypothetical AV faces a two-alternative forced choice of whom to hit or save—e.g., a driver vs. a pedestrian; a homeless man vs. a skilled workman (19). The studies ask people to choose on the AV's behalf, and then they aggregate the choices to assemble a "global preference" scale, which they suggest should be considered in AV policy. One implication of this crowdsourcing work, whether originally intended or interpreted by others, is that AVs should be programmed with moral rules for solving these dilemmas (19–24), an idea that has received overwhelming attention from the media (25–32). Reactions from the academic community have been more mixed (33–38), and the few reactions from the AV community have been largely dismissive (39–42).

Driverless Dilemmas Reflect a Consensus That AVs Should Generally Conform to Commonsense Social Expectations. Are driverless dilemmas a good way to think about AV ethics? In *SI Appendix*, Table S1, we briefly review the main affirmative and negative answers that have been offered to the question raised above (19, 34–36, 39, 40, 43). While we will not try to decisively settle these arguments, we think that they highlight three important principles in evaluating driving behavior. First, they capture the notion that AVs should generally conform to certain public expectations, because they must behave with what people tend to call common sense. A person may think that a driver has common sense when that driver makes what appear to be optimal, harm-reducing choices (e.g., maneuvers or speed). Even mundane-seeming deviations from common sense (e.g., how an AV avoids obstacles in the road) may put others at risk, so designing these behaviors may rely, in part, on ethical inputs that arbitrate between multiple, viable engineering solutions. On this view, people may be especially interested in dilemma scenarios because they involve difficult trade-offs between various maneuvers in their risk of harm to others, such that seeing how the AV solves these challenges might suggest whether it is likely to exhibit common sense more generally.

Second, within the limits of their scenario abstraction, driverless dilemmas are based on a scientific way to look at AV behavior to the extent that they take a falsifiable behavioral approach: specifically, to determine whether the AV is behaving properly from external observations of its behavior, rather than from speculations about the underlying "mental processes" that prompted the behavior. In engineering, this black-box approach (not to be confused with event data recorders of acceleration, speed, etc.; also referred to as "black boxes"), or "performance standard" in law, is one extreme on a shades-of-gray continuum. The opposite would be a clear-box approach, which would entail inspecting the source code of all the AV's components—and, under a "design standard" in law, regulating them directly. From a standardized testing point of view, a clear-box approach may be less effective for AVs. It is much harder for testers outside of the company to scrutinize, regulate, and gain access to the software itself, which is extremely hard to understand for nonpractitioners, and prone to mischaracterization if simplified too much. Furthermore, the AV's overall behavior is an emergent property of various software components that cannot easily be reduced to any single component. In short, as in the case of tort law and Turing tests (44), it seems reasonable to specify and assess AV behavior in part by measuring that behavior itself, rather than by attempting to decipher partially knowable components and inaccessible mental states. This analysis is itself imperfect, as tragedies like the two Boeing 737 MAX crashes have shown (45), but it does suggest why falsifiable behavioral testing will be appealing in practice. However, ultimately, as we explain in the final section, such testing is just one piece of the overall process of integrating ethics into the design and evaluation of AVs.

Third, the popularity of these dilemmas suggests that they are simple and easy to think about. The explanation and regulation of more practical and relevant issues might also aspire to this kind of public accessibility.

A Pragmatic Implication: Test for Commonsense Behavior

These positive features can form part of more practical assessments for the AV industry that directly test for commonsense behavior. These tests should serve two overarching, complementary functions: 1) Explanatory role. The qualitative analysis of the scenarios teaches people that there are trade-offs in consequences depending on the ethical principles that we program into AVs. 2) Behavior assessment. The scenarios empirically test whether the AV produces the correct set of behaviors across a variety of practical situations.

We note that designing the testing machinery is a different problem than choosing the "right answers" to the tests. The first is a technical problem, whereas the second is a social and regulatory problem that ultimately relies on ethical inputs too. After first saying a little more about what we mean by commonsense behavior, we discuss both problems: how to design more technically pragmatic tests, and how to revise current thinking on their right answers.

What It Means for an AV to Drive with Common Sense. There have been various ideas for how to formulate autonomous machines that have common sense (46, 47). In the case of automated driving, the AV must execute a set of control commands (e.g., steering, brake, throttle, gear position) that achieve the desired transportation goal (e.g., delivering occupants or goods from point A to B) (6, 48). More than one engineering solution is probably viable. For instance, in the optimal control method (49), the AV minimizes the error between the path it is taking and the path it is planning by making adjustments (e.g., steering adjustments) at various time intervals. Behaving optimally under these circumstances is ethically relevant because the constraints on this path include not hurting others or putting them at risk for harm.

However, commonsense driving also entails more. For instance, an AV that makes too many adjustments spoils the smoothness of the ride. Engineers must deal with this issue by, for instance, placing an additional cost on changing steering every time interval, thereby raising the margin by which the AV may veer from its planned path before it is permitted to execute its next adjustment (48).

These tensions show how achieving commonsense driving in AVs necessarily implicates the broad and interdisciplinary field of human factors. Ultimately, passengers and other road users may expect AVs to behave in a way that is safe, predictable, reasonable, uniform, comfortable, and explainable (SPRUCE). These concepts are naturally interrelated, but also capture distinct components of driving behavior, for which we provide high-level definitions here:

- (S) Safe. The AV does not harm others or put others at unreasonable risk of harm. As with inferences about the moral character of people (50), determining whether an AV is safe informs predictions about how it will generally behave in both the near future and in new situations, regardless of the details.
- (P) Predictable. The AV's maneuvers can be anticipated from past behavior.
- (R) Reasonable. The AV's behaviors do not offend notions of logic or justice and generally accord with human intuition (51, 52).
- (U) Uniform. The AV behaves consistently by treating seemingly like situations alike.
- (C) Comfortable. The AV drives in a manner that is physically and psychologically smooth for its passengers and for other road users.
- (E) Explainable. The AV's actions fit an accessible narrative of cause and effect, action and reaction, and stimulus and response (53, 54). [This notion of popular explainability complements the technical concept of "explainable AI," in which algorithmic decision making is transparent and traceable at a technical level (55, 56).]

However, these behavioral expectations represent a rough floor rather than ceiling for AVs. At the end of the day, AVs should be better than human drivers. This may sometimes require AVs to behave in internally consistent ways that may nonetheless seem inconsistent to human observers. However, this in turn requires the AV to acclimate the humans with which it interacts. While humans may seem predictable to each other, they may also routinely behave suboptimally. In part, AVs may seem less predictable to humans for the very reason that they are objectively more predictable, unwaveringly obeying their programs and the rules of the road (57). Is this not the sort of standard we are already meant to require of drivers? In other cases, AVs may need to acclimatize to humans; e.g., an AV that slows down every time it detects a crosswalk could lead a pedestrian to infer that it is communicating its intent to stop, even if it has not actually detected her (57). Safety can be improved by making the AV more human-like in certain ways (such as adding a more explicit signal that it will stop) rather than others (e.g., speeding through crosswalks unless doing so exceeds a high-risk threshold). More generally, the question should not be "who should acclimatize to whom?" but "what system is the safest?," and the best solutions might involve reconceptualizing entire driving systems rather than just acclimatizing drivers. AVs should set new norms that push everyone to improve (58).

What It Means for People to Be Persuaded That an AV Drives with Common Sense. Whether an evaluator is persuaded that an AV has common sense is a judgment that, itself, is reliant on mental processes that we might call common sense. Psychologically speaking, common sense is thought to arise from complex-but-fast processes that support our intuitions about a range of phenomena pertaining to both the physical domain (e.g., space, time, objects, persistence) and the social domain (e.g., mental states, motivations, moral character). As just one example, if a vehicle enters and emerges from occlusion, we understand when it is the same vehicle rather than another vehicle altogether, and we can anticipate when it will emerge before it does (59, 60). Most people agree on basic intuitions pertaining to these domains, even if they do not know how they came to have these intuitions, hence the "common" in common sense. Common sense is typically contrasted with slower, more deliberative and effortful processing, as when multiplying large numbers or choosing among several similar options (61).

Notably, people are highly sensitive to whether other agents behave optimally given their constraints. This is an assessment that even infants are capable of (62, 63), that occurs across various domains of human cognition (54, 64–67), and that affects moral intuitions in adults about whether an agent is blameworthy or morally wrong (53). People find suboptimal choices more difficult to explain, and this intuition leads them to assign harsher moral judgments accordingly (53). Thus, AVs that make suboptimal maneuvers are likely to seem inexplicable too, leading to the impression that they are morally faulty and culpable for these violations. In this way, an evaluator's own common sense can guide their assessment of whether an AV has common sense, a sign of whether the vehicle is ethically roadworthy.

Testing an AV's Ability to Drive with Common Sense. Humans often deal with new situations beyond the formal driving test by using common sense (provided they have the time to do so). However, it is not appropriate to assume that AVs will be similarly reasonable. If we only test their obedience to secondary road rules under ordinary conditions, then broader safe behavior is not guaranteed. Driving common sense is at play in various choices, such as deciding what to do when a traffic light fails, or ascertaining whether a pedestrian is about to cross the road. And while it is easy to stipulate that an AV should never fail, getting there requires looking beyond redundancy (such as increasing the number of sensors that detect the same information) to resiliency (creating a behavioral system that can solve problems in various driving settings), such as executing more predictable behaviors that minimize risky encounters with humans and prevent system failures from escalating into catastrophes.

In order to develop more pragmatic tests of whether an AV behaves safely, we think the above ideas can inspire "edge-case" testing of "driving common sense" in AVs—that is, testing whether an AV's behavior is broadly safe, predictable, reasonable, uniform, comfortable, and explainable. We expect such an AV to minimize harm or the risk of harm across a series of "low-stakes" and "high-stakes" scenarios. By low-stakes scenarios, we mean high-probability situations wherein no outcome results in death or injury, but in which the vehicle's decisions still require arbitrating between more vs. less risky paths. By high-stakes scenarios, or edge cases, we mean low-probability, dilemma-like scenarios, wherein the vehicle's possible paths are mathematically constrained so that a collision is inevitable (48, 68). As in any safety evaluation of edge cases, these tests must cover a range of scenarios that are not typically encountered during a standard test for human drivers, but which may nonetheless arise on real roads, such as navigating around large debris, safely overtaking long trucks on a two-way road, making way for an emergency vehicle, or being manually directed by a traffic officer (69). The scenarios can vary along multiple dimensions, such as prevalence and ethical relevance, and be tested both in simulation and on real roads (except for high-stakes scenarios). The tests should also be adaptive, changing to accommodate the newest systems as they become available, as when software updates are deployed over the air to AVs.

Such edge-case testing would be beneficial to several stakeholders: 1) manufacturers, as internal tools for development; 2) regulators, for regulation of the technologies; 3) insurers, who need to quantify the risk of ensuring a technology; and 4) the public, who seek assurance that the vehicles behave ethically. If an AV passes a range of scenarios that involve driving common sense, then we may be more confident that it is operating based on algorithms that are sufficiently flexible and robust to minimize the risk of harm on real roads. In developing such tests, incorporating the following six factors is beneficial (Fig. 1):

1) Test common sense as an integral part of road rules testing. A problem of behavior specification for AVs is the interpretation of ambiguous laws such as "right of way," which need to be encoded into a machine-readable definition. Adding complication, new driving situations may arise in which road rules conflict among themselves and with broader ethical principles, as in Asimov's laws (70). Such hierarchies are often already encoded in the law. For example, the rule to follow the right of way entitles no one to gratuitously hit another. In fact, law often speaks only of an obligation to yield the right of way, so that following the rule itself comes second to its main goal of avoiding harm (71). Another example is a police officer who instructs a driver to do something, even if it means breaking a usual rule like crossing a solid white line (72).
2) Test basic behavioral competencies alongside high-stakes scenarios. Low-stakes scenarios that test for basic behavioral competencies (73) may help to illuminate technical issues without the emotional distraction of rarer high-stakes scenarios (48, 68, 74). Once the trade-offs of low-stakes scenarios are understood, engineers can slowly scale up the stakes to be included in their simulations. As one example, driverless dilemmas grapple with the edge case of minimizing collisions with other road users in the catastrophic scenario where a collision is inevitable. A low-stakes variant of this problem is how to avoid road debris. An AV with common sense will try to avoid coming to an abrupt standstill, maneuvering around the obstacle while maintaining a "safe distance" from both the obstacle and other oncoming drivers (68). Such simple obstacle avoidance still has ethical implications and will probably need to be solved before the AV has a hope of solving more complicated, high-stakes variants in which multiple obstacles are constraining its path.
3) Acknowledge the trade-offs of multiple metrics. Driving entails resolving differences between conflicting objectives. For instance, in the previous example, the AV must balance safety (e.g., avoid a collision) and compliance (e.g., do not cross the solid white line) (68). In fact, the paramount command in driving law in the United States is to avoid harm, so compliance with the law may sometimes require, rather than prohibit, nominally violating a road rule in order to prevent harm, e.g., driving on the shoulder of the road in order to prevent an otherwise unavoidable collision. Existing law assumes that human drivers can navigate such trade-offs (72, 75–78).
4) Add notions of uncertainty. In practice, it is not guaranteed that taking a given action (e.g., swerving to the right) will produce a specific outcome in a deterministic fashion (e.g., certainly avoiding a collision with a neighboring vehicle). Rather, there is some probability that the outcome will occur, contingent on the action. A prevalent challenge is determining what threshold a given probability must surpass before the AV should execute an action, given that it must also estimate its certainty in this probability (74, 79). More broadly, analyses must move beyond harm to assessing risk of harm, and then from risk to assessing confidence about that risk.
5) Add the notion of states of information. Sometimes it is not possible to identify what is outside of the vehicle, increasing the chance of false positives or negatives (79, 80), e.g., assuming there is a pedestrian in front of the vehicle when

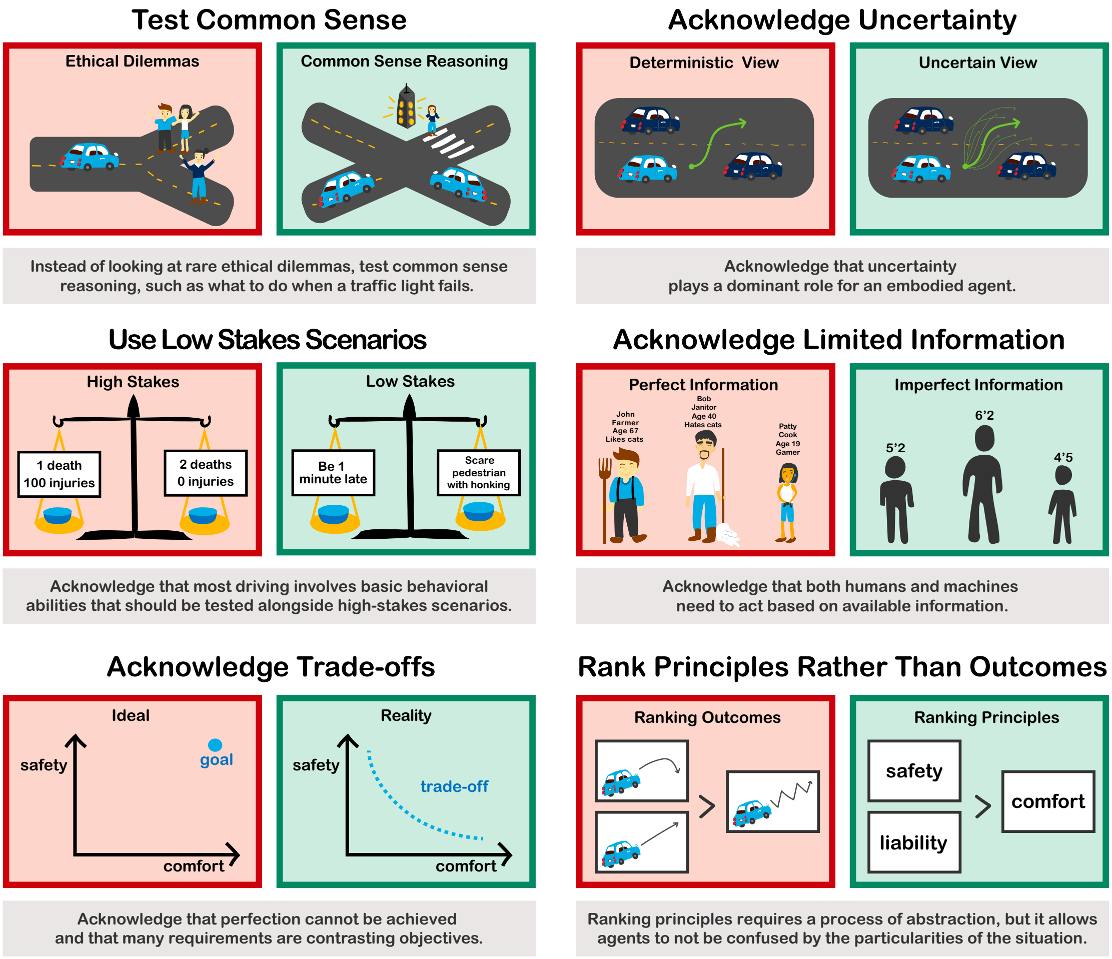
Test Common Sense
Ethical Dilemmas
Common Sense Reasoning
Instead of looking at rare ethical dilemmas, test common sense reasoning, such as what to do when a traffic light fails.
Acknowledge Uncertainty
Deterministic View
Uncertain View
Acknowledge that uncertainty plays a dominant role for an embodied agent.
Use Low Stakes Scenarios
High Stakes
1 death 100 injuries
2 deaths 0 injuries
Low Stakes
Be 1 minute late
Scare pedestrian with honking
Acknowledge that most driving involves basic behavioral abilities that should be tested alongside high-stakes scenarios.
Acknowledge Limited Information
Perfect Information
John Farmer Age 67 Likes cats
Bob Janitor Age 40 Hates cats
Patty Cook Age 19 Gamer
Imperfect Information
5'2
6'2
4'5
Acknowledge that both humans and machines need to act based on available information.
Acknowledge Trade-offs
Ideal
safety
goal
comfort
Reality
safety
trade-off
comfort
Acknowledge that perfection cannot be achieved and that many requirements are contrasting objectives.
Rank Principles Rather Than Outcomes
Ranking Outcomes
>
Ranking Principles
safety
liability
>
comfort
Ranking principles requires a process of abstraction, but it allows agents to not be confused by the particularities of the situation.

Fig. 1. From driverless dilemmas (red) to more practical recommendations for testing an AV's ability to drive with common sense (green).

no one is there. Modern AV developers factor estimates of these probabilities into AV decision-making by setting thresholds. Adding complexity, some existing algorithms may not accurately identify unanticipated objects that are otherwise perfectly identifiable to humans. For example, we might reasonably assume that a human driver who correctly identifies and appropriately responds to a stop sign at noon will also correctly identify and appropriately respond to a yield sign at dawn. In contrast, an algorithm that has never encountered a yield sign at dawn may respond in an unexpected way. Indeed, the first person to die in a collision with an AV undergoing testing was a homeless woman pushing a bike with bags across a suburban street in the evening (79).

6) Rank principles rather than outcomes. One problem with ranking ethically relevant choices based on their resulting outcomes is that many possible outcomes can result from any given choice. Therefore, in practice it may be more tractable to rank principles, e.g., avoid hitting humans > follow road rules > take the fastest route (8, 48, 68).

How to Obtain the Right Answers to Commonsense Tests. So far, we have explained how to make more realistic and useful commonsense diagnostics for AVs. However, there is another issue that we see as orthogonal: What are the right answers to high-stakes, dilemma scenarios? We suggest five ways of improving current thinking on this issue.

1) Do not assume that human driving is a gold standard. Although humans can be skilled drivers, it is important to remember that existing road fatality rates occur because of the collective failure of road operators, vehicle manufacturers, and road users (1, 2). Existing human driving patterns can be quantified along several metrics to understand how people ordinarily adhere to otherwise vague notions like right of way (81), and these driving patterns should be considered when deciding how to program AVs. However, we should carefully choose what subset of humans to mimic, if any at all. Similar lessons have been learned in older fields like clinical decision-making, where statistical models are already known to outperform trained clinicians in certain cases (82, 83). Law, for its part, expects what is reasonable under the circumstances—not merely what would be reasonable for a human (4, 51). A related issue is how to facilitate safe interactions between AVs and human drivers, who tend to interpret road rules loosely rather than literally, e.g., yielding at stop signs, ignoring pedestrians on a crosswalk, speeding in neighborhoods, driving in the bicycle lane. The safest approach may be to raise the bar for everyone, by more strictly enforcing road rules for existing human drivers, rather than engineering AVs that drive like humans (58).
2) Reconsider existing notions of control. Some variants of driverless dilemmas ask people to consider cases in which human drivers have varying degrees of control (5), then they

ask people to rate who is more blameworthy for harmful outcomes—the human, AV, or manufacturer (84, 85). However, ultimately it may become less safe to allow humans to switch dynamic control between themselves and AVs in the first place, than to relinquish dynamic control to the AV (85, 86). Safer roads may require getting used to this idea, rather than settling for suboptimal outcomes just to preserve human authority and discretion in a few cases (4). Humans already surrender control to some other technologies such as elevators, which effectively trap us during malfunction until the engineers arrive (4). However, relinquishing some kinds of control will likely depend on more widespread understanding of and trust in how AVs behave (4, 48).

3) Do not blindly assume that AVs will learn everything from the data. Some behaviors might be learned for low-stakes scenarios, but it is unlikely that AVs can learn how to deal with scarce edge cases. Therefore, we cannot simply assume that best behavior in such edge cases will be "taken care of," as we do (for better or worse) with human drivers who pass standard driving tests. Rather, as in any unit testing setting, we may need to test for potential boundary conditions on the AV's behavior in order to ensure that it behaves robustly across a range of parameters. An AV that behaves well at the extremes is more likely to also do so in intermediate scenarios; yet the same does not apply in reverse, since extreme edge cases are more likely to involve additional factors that do not arise in intermediate scenarios. As an example, the rule to never reverse while stopped on an incline is generally effective yet must be violated in the edge case where the vehicle in front of the AV is rolling backward and there is sufficient space for the AV to reverse. Testing for this edge case can expose a faulty algorithm that would otherwise pass a standard test. More to the point, we can test many of the intermediate scenarios as well.

4) Allow for local customization. Since de facto rules of the road already vary across countries, we should not expect there to be a unique, globally consistent AV common sense—even if AV behavior can be formally specified in a way that might seem to make cultural idiosyncrasies irrelevant. Rather, we may have to accept that the right answers to government-administered or mandated driving tests of common sense will change from jurisdiction to jurisdiction. As an example, a recent report from the German Ministry of Transportation prescribed: "In the event of unavoidable accident situations, any distinction between individuals based on personal features (age, gender, physical, or mental constitution) is impermissible" (87). Other countries may conclude differently, necessitating algorithms that anticipate or flexibly adjust to local driving rules. AV design will likely be jurisdictionally dependent in many other ways as well, from rules of the road to traffic control devices to traffic patterns.

5) Go beyond binary choices and right–wrong evaluations. In real-world situations, there is a continuum of possible actions and corresponding scores for ranking them. Maybe hardest for the public to understand is that there is no such thing as perfect technology. Rather, there are various trade-offs between benefits and risks, just as in medicine, e.g., the trade-off between safer but more expensive vaccines vs. vaccines that cause more adverse events but cost less and hence may be more widely available (88). Furthermore, it is not only the behavior of the AV itself that is ethically relevant but also that of the transport infrastructure that a government can build in order to minimize the frequency and magnitude of risky traffic scenarios, e.g., whereas crossing a street is often dangerous and inefficient for pedestrians today, future roads could be built to ensure that pedestrians can safely cross almost anywhere on the street, safely facilitating their natural inclination to do so.

What Would a Decision Process for Defining the Tests and Their Answers Look Like? How should we create these tests and define their right answers? The idea of testing for common sense raises a number of questions. Should the tests be a public process? Should they be developed in conjunction with all stakeholders, or just some, e.g., the regulator, or the company? Who has the burden of verification? Should the details of the testing process be allowed to vary from country to country? Should human data, e.g., in the form of judgments or behavior, play any part in informing development of the tests?

Although these questions extend beyond the scope of this article, it is worth making a few general points. First, commonsense testing will not be limited to the deployment stage but will exist in all the spheres we have identified (internal development, insurance, regulation, publication evaluation, litigation, etc.). Second, it is likely that the exact form that these tests take will look different in each of these domains. Third, the design elements of these tests will themselves involve ethical choices, e.g., whether to assume danger, or to assume safety. Fourth, these ethical issues are not unique to AVs, cf. pharmaceuticals (89, 90).

The Road Ahead: Ethics as a Marriage, Not a Wedding

Our analysis has led to two open questions: First, how should we evolve beyond driverless dilemmas for the sake of realism and relevance to actual AVs? Second, what is the best process to obtain the right answers to a broader range of scenarios? We provided some first answers to these questions, offering a view of the long road ahead.

However, it is not enough to stop there. Once-off assessments of human or machine driving with a single licensing test—even one that assesses common sense behavior—is far from a gold standard for safety. A good ethical testing framework for AVs should ensure the safety of the vehicles at every stage of development, deployment, and operation, and so requires a holistic approach that emphasizes process (4). This means focusing just as much on the company as on the product, and scrutinizing factors such as "corporate governance, design philosophy, hiring and supervision, evaluation and integration of standards, monitoring and updating, communication and disclosure, and planning for eventual obsolescence" (4). Companies should also carry the public-facing responsibility of saying "what they are doing, why they think it is reasonably safe, and why we should believe them" (4). For an AV manufacturer, this includes not only developing and demonstrating the sort of tests we have outlined here, but also anything else that is needed to make its safety case, as well as being upfront about what is hard and did not work (91–96). If companies can legitimately convince consumers that they are transparent, vigilant, and continuously improving, then consumers will reasonably trust them. Our analysis concludes that at least one key part of earning this trust will be convincing the public that their AVs behave with common sense.

In regulating companies, a certain amount of "fuzziness" in the law may be beneficial, since overly specific rules may be unnecessarily restrictive (especially if the technology is not yet ripe), and the question for companies should be "is this safe," not "is this compliant," e.g., if there is a defect in a safety system that the company is not strictly obligated to fix, law should still vaguely obligate them to fix it (4). At the same time, overly vague "declarations of ethical principles," as in the case of Germany, may also add to the confusion. [Notably, much of the German report also focused on high-stakes dilemma scenarios, including two ethical rules devoted to this topic (97).] Instead, regulators should directly collaborate with (without deferring to) developers to define principles that will not cause needless hassles later on. Countries like Singapore provide an admirable example, since they have engaged directly with AV manufacturers and engineers to iteratively refine AV regulation in sufficient technical detail that it is actionable and adherable—likely driven by their practical need for an AV vehicle-sharing solution to growing traffic congestion on the small island (81). Regulators should learn about autonomy and work with manufacturers to develop facilitative legal regimes that prioritize safety over mere compliance with the rules.

There are also various steps that other stakeholders can take to increase transparency and vigilance. Engineers should not ignore ethical questions that arise from their designs, nor try to "solve" them by unilaterally, opaquely, and informally implementing their own best judgments. Testing entities should release more comprehensive reports that detail the limitations of their tests of vehicle safety, rather than only summarizing vehicle achievements. Insurance companies should be more forthcoming in sharing their risk knowledge and pricing methodologies. Science communicators and social scientists should prioritize articles on serious, public health issues surrounding AVs, instead of click-bait headlines (98). Social scientists interested in policy should study the public perception challenges remaining in the path toward fully safe AVs, while focusing on questions that are practical and relevant. They can also work with engineers to collect reasonable data on human driving patterns, which may inform AV choices when existing road rules are underspecified. All stakeholders can infuse ethics in their actions and in their expectations.

Data Availability. There are no data underlying this work.

Acknowledgments

We thank Ed Niedermeyer, Marcus Baram, Keith Shaw, Noah Goodall, Steven Pinker, Peter DeScioli, and George A. Alvarez for helpful discussion and Patcharapon Thammathorn for designing the figure.

1 R. Nader, *Unsafe at Any Speed: The Designed-In Dangers of the American Automobile* (Grossman, New York, 1965).

2 B. Welle *et al.*, "Sustainable and safe: A vision and guidance for zero road deaths" (World Resources Institute, 2018). https://files.wri.org/s3fs-public/sustainable-safe.pdf. Accessed 1 June 2020.

3 S. Singh, *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey* (National Highway Traffic Safety Administration, 2015).

4 B. W. Smith, "Ethics of artificial intelligence in transport" in *The Oxford Handbook of Ethics of Artificial Intelligence*, M. Dubber, F. Pasquale, S. Das, Eds. (Oxford University Press, 2020), pp. 667–683.

5 SO-RAVS Committee, Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. *SAE Standard J.* **3016**, 1–16 (2014).

6 J. C. Gerdes, S. M. Thornton, J. Millar, "Designing automated vehicles around human values" in *Automated Vehicles Symposium*, G. Meyer, S. Beiker, Eds. (Springer, 2019), pp. 39–48.

7 B. W. Smith, "Lawyers and engineers should speak the same robot language" in *Robot Law*, R. Calo, A. M. Froomkin, I. Kerr, Eds. (Edward Elgar Publishing, 2016), pp. 78–101.

8 B. W. Smith, "Regulation and the risk of inaction" in *Autonomes Fahren*, M. Maurer, J. C. Gerdes, B. Lenz, H. Winner, Eds. (Springer, 2015), pp. 571–587.

9 A. Tversky, C. R. Fox, Weighing risk and uncertainty. *Psychol. Rev.* **102**, 269–283 (1995).

10 A. Tversky, D. Kahneman, Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992).

11 R. Hotten, Volkswagen: The scandal explained. BBC News, 10 December 2015. https://www.bbc.com/news/business-34324772. Accessed 1 June 2020.

12 P. Foot, The problem of abortion and the doctrine of double effect. *Oxford Rev.* **5**, 5–15 (1967).

13 S. Kagan, "Solving the trolley problem" in *The Trolley Problem Mysteries*, F. M. Kamm, E. Rakowski, Eds. (Oxford University Press, New York, 2016), pp. 151–166.

14 J. Greene, "Solving the trolley problem" in *A Companion to Experimental Philosophy*, J. Sytsma, W. Buckwalter, Eds. (Wiley, 2016), pp. 173–189.

15 J. J. Thomson, The trolley problem. *Yale Law J.* **94**, 1395–1415 (1984).

16 J. De Freitas, P. DeScioli, J. Nemirow, M. Massenkoff, S. Pinker, Kill or die: Moral judgment alters linguistic coding of causality. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 1173–1182 (2017).

17 G. Marcus, Moral machines. *The New Yorker*, 24 November 2012. https://www.newyorker.com/news/news-desk/moral-machines. Accessed 1 June 2020.

18 B. W. Smith, *Driving at Perfection* (The Center for Internet and Society at Stanford Law School, 2012).

19 E. Awad *et al.*, The moral machine experiment. *Nature* **563**, 59–64 (2018).

20 J.-F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).

21 P. Lin, "Why ethics matters for autonomous cars" in *Autonomous Driving*, M. Maurer, J. Gerdes, B. Lenz, H. Winner, Eds. (Springer, Berlin, 2016), pp. 69–85.

22 J. D. Greene, ETHICS. Our driverless dilemma. *Science* **352**, 1514–1515 (2016).

23 J. Gogoll, J. F. Müller, Autonomous cars: In favor of a mandatory ethics setting. *Sci. Eng. Ethics* **23**, 681–700 (2017).

24 R. Noothigattu *et al.*, "A voting-based system for ethical decision making" in *Thirty-Second AAAI Conference on Artificial Intelligence* (Association for the Advancement of Artificial Intelligence, 2018), pp. 1587–1594.

25 J. Donde, Self-driving cars will kill people. who decides who dies? *Wired*, 9 September 2017. https://www.wired.com/story/self-driving-cars-will-kill-people-who-decides-who-dies/. Accessed 1 June 2020.

26 D. Edmonds, Cars without drivers still need a moral compass, but what kind? *The Guardian*, 14 November 2018. https://www.theguardian.com/commentisfree/2018/nov/14/cars-drivers-ethical-dilemmas-machines. Accessed 1 June 2020.

27 C. Y. Johnson, Self-driving cars will have to decide who should live and who should die: Here's who humans would kill. *The Washington Post*, 24 October 2018. https://www.washingtonpost.com/science/2018/10/24/self-driving-cars-will-have-decide-who-should-live-who-should-die-heres-who-humans-would-kill/. Accessed 1 June 2020.

28 C. A. Lester, A study on driverless-car ethics offers a troubling look into our values. *The New Yorker*, 24 January 2019. https://www.newyorker.com/science/elements/a-study-on-driverless-car-ethics-offers-a-troubling-look-into-our-values. Accessed 1 June 2020.

29 P. Lin, The ethics of autonomous cars. *Atlantic*, **8** October 2013. https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. Accessed 1 June 2020.0276-9077.

30 J. Markoff, Should your driverless car hit a pedestrian to save your life? *New York Times*, 23 June 2016. https://www.nytimes.com/2016/06/24/technology/should-your-driverless-car-hit-a-pedestrian-to-save-your-life.html?auth=linked-google. Accessed 1 June 2020.

31 P. Nowak, The ethical dilemmas of self-driving cars. *The Globe and Mail*, 2 February 2018. https://www.theglobeandmail.com/globe-drive/culture/technology/the-ethical-dilemmas-of-self-drivingcars/article37803470/. Accessed 1 June 2020.

32 A. Shariff, I. Rahwan, J.-F. Bonnefon, Whose life should your car save? *New York Times*, 3 November 2016. https://www.nytimes.com/2016/11/06/opinion/sunday/whose-life-should-your-car-save.html. Accessed 1 June 2020.
33 B. Dewitt, B. Fischhoff, N. E. Sahlin, "Moral machine" experiment is no basis for policymaking. *Nature* **567**, 31 (2019).
34 G. Keeling, Why trolley problems matter for the ethics of automated vehicles. *Sci. Eng. Ethics* **26**, 293–307 (2020).
35 J. Himmelreich, Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory Moral Pract.* **21**, 669–684 (2018).
36 N. J. Goodall, Away from trolley problems and toward risk management. *Appl. Artif. Intell.* **30**, 810–821 (2016).
37 Y. E. Bigman, K. Gray, Life and death decisions of autonomous vehicles. *Nature* **579**, E1–E2 (2020).
38 J. De Freitas, M. Cikara, Deliberately prejudiced self-driving cars elicit the most outrage. *Cognition* **208**, 104555 (2021).
39 E. Olson, *Trolley Folly* (Medium, 2018).
40 J. De Freitas, S. E. Anthony, A. Censi, G. A. Alvarez, Doubting driverless dilemmas. *Perspect. Psychol. Sci.* **15**, 1284–1288 (2020).
41 K. Iagnemma, Why we have the ethics of self-driving cars all wrong. World Economic Forum, 21 January 2018. https://www.weforum.org/agenda/2018/01/why-we-have-the-ethics-of-self-driving-cars-all-wrong/. Accessed 1 June 2020.
42 B. W. Smith, Slow down that runaway ethical trolley. CIS Blog. (2016), 12 January 2015. https://cyberlaw.stanford.edu/blog/2015/01/slow-down-runaway-ethical-trolley. Accessed 1 June 2020.
43 S. Nyholm, J. Smids, The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory Moral Pract.* **19**, 1275–1289 (2016).
44 A. M. Turing, Computing machinery and intelligence. *Mind* **LIX**, 23–65 (1950).
45 J. Herkert, J. Borenstein, K. Miller, The Boeing 737 MAX: Lessons for engineering ethics. *Sci. Eng. Ethics* **26**, 2957–2974 (2020).
46 J. McCarthy, *Programs with Common Sense* (RLE and MIT Computation Center, 1960).
47 M. Minsky, Commonsense-based interfaces. *Commun. ACM* **43**, 66–73 (2000).
48 J. C. Gerdes, S. M. Thornton, "Implementable ethics for autonomous vehicles" in *Autonomes Fahren*, M. Maurer, J. Gerdes, B. Lenz, H. Winner, Eds. (Springer, 2015), pp. 87–102.
49 V. G. Boltyanskiy, R. V. Gamkrelidze, L. S. Pontryagin, On the theory of optimal processes. *Dokl. Akad. Nauk SSSR* **110**, 7–10 (1956).
50 J. De Freitas, P. DeScioli, K. A. Thomas, S. Pinker, Maimonides' ladder: States of mutual knowledge and the perception of charitability. *J. Exp. Psychol. Gen.* **148**, 158–173 (2019).
51 B. W. Smith, *Automated driving and product liability* (St. L. Rev, Mich., 2017), vol 1, https://digitalcommons.law.msu.edu/lr/vol2017/iss1/1. Accessed 1 June 2020.
52 K. P. Tobia, How people judge what is reasonable. *Ala. Law Rev.* **70**, 293 (2018).
53 J. De Freitas, S. G. Johnson, Optimality bias in moral judgment. *J. Exp. Soc. Psychol.* **79**, 149–163 (2018).
54 S. G. Johnson, L. J. Rips, Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognit. Psychol.* **77**, 42–76 (2015).
55 W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv:1708.08296 (2017).
56 D. Gunning, Explainable artificial intelligence (DARPA, 2017). https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf. Accessed 1 June 2020.
57 H. Surden, M.-A. Williams, Technological opacity, predictability, and self-driving cars. *Cardozo Law Rev.* **38**, 121 (2016).
58 B. W. Smith, How governments can promote automated driving. *N. M. Law Rev.* **47**, 99 (2017).
59 J. De Freitas, N. E. Myers, A. C. Nobre, Tracking the changing feature of a moving object. *J. Vis.* **16**, 22 (2016).
60 A. D. Makin, E. Poliakoff, W. El-Deredy, Tracking visible and occluded targets: Changes in event related potentials during motion extrapolation. *Neuropsychologia* **47**, 1128–1137 (2009).
61 D. Kahneman, *Thinking, Fast and Slow* (Macmillan, 2011).
62 G. Csibra, G. Gergely, S. Bíró, O. Koós, M. Brockbank, Goal attribution without agency cues: The perception of "pure reason" in infancy. *Cognition* **72**, 237–267 (1999).
63 G. Gergely, Z. Nádasdy, G. Csibra, S. Bíró, Taking the intentional stance at 12 months of age. *Cognition* **56**, 165–193 (1995).
64 D. C. Dennett, *The Intentional Stance* (MIT Press, 1989).
65 T. Gao, B. J. Scholl, Chasing vs. stalking: Interrupting the perception of animacy. *J. Exp. Psychol. Hum. Percept. Perform.* **37**, 669–684 (2011).
66 D. Davidson, "Truth and meaning" in *Philosophy, Language, and Artificial Intelligence*, J. Kulas, J. H. Fetzer, T. L. Rankin, Eds. (Springer, 1967), pp. 93–111.
67 H. P. Grice, *Studies in the Way of Words* (Harvard University Press, 1989).
68 A. Censi *et al.*, "Liability, ethics, and culture-aware behavior specification using rulebooks" in *2019 International Conference on Robotics and Automation (ICRA)* (IEEE, 2019), pp. 8536–8542.
69 N. H. T. S. Administration, *Automated Driving Systems 2.0: A Vision for Safety* (US Department of Transportation, 2017).
70 I. Asimov, Runaround. *Astounding Sci. Fiction* **29**, 94–103 (1942).
71 South Carolina Code of Laws, 56-5 §§ 2310, 3230 (1962).
72 B. W. Smith, Automated vehicles are probably legal in the United States. *Tex. A&M L. Rev.* **1**, 411 (2013).
73 Program UoCP, *Peer Review of Behavioral Competencies for AVs* (University of California PATH Program, 2016).
74 B. W. Smith, The Trolley and the Pinto: Cost-benefit analysis in automated driving and other cyber-physical systems. *Tex. A&M L. Rev.* **4**, 197 (2016).
75 Tex. Penal Code Ann., § 9.22(d) (1994).
76 Judicial Council of California Criminal Jury Instructions, § 3403 (2018).
77 Com. v. Livington, 70 Mass. App. Ct. 745, 749, 877 N.E.2d 255 (2007).
78 State v. Riedl, 15 Kan. App. 2d 326, 331, 807 P.2d 697 (1991).
79 National Transportation Safety Board, *Preliminary report, highway, HWY18mh010* (National Transportation Safety Board, 2018), pp. 1–4.
80 D. M. Green, J. A. Swets, *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
81 Singapore Standards Council, *TR 68: Technical Reference for Autonomous Vehicles* (Singapore Standards Council, 2019).
82 A. S. Elstein, A. Schwartz, Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ* **324**, 729–732 (2002).
83 D. Wedding, Clinical and statistical prediction in neuropsychology. *Clin. Neuropsychol.* **5**, 49–55 (1983).
84 E. Awad *et al.*, Drivers are blamed more than their automated cars when both make mistakes. *Nat. Hum. Behav.* **4**, 134–143 (2020).
85 R. M. McManus, A. M. Rutchick, Autonomous vehicles and the attribution of moral responsibility. *Soc. Psychol. Personal. Sci.* **10**, 345–352 (2019).
86 B. W. Smith, Controlling humans and machines. *Temp. Intl. Comp. LJ* **30**, 167 (2016).
87 BMVI, "Ethics Commission. Automated and connected driving" (Federal Ministry of Transport and Digital Infrastructure, 2017).
88 M. A. Miller, R. W. Sutter, P. M. Strebel, S. C. Hadler, Cost-effectiveness of incorporating inactivated poliovirus vaccine into the routine childhood immunization schedule. *JAMA* **276**, 967–971 (1996).
89 M. Angell, The ethics of clinical research in the Third World. *N. Engl. J. Med.* **337**, 847–849 (1997).
90 K. Dickersin, D. Rennie, Registering clinical trials. *JAMA* **290**, 516–523 (2003).
91 National Conference of Commissioners on Uniform State Laws, Uniform Automated Operation of Vehicles Act. Uniform Law Commission (2019). https://www.uniformlaws.org/viewdocument/final-act-with-comments-133?CommunityKey=4e70cf8e-a3f4-4c55-9d27-fb3e2ab241d6&tab=librarydocuments. Accessed 1 June 2020.
92 National Highway Traffic Safety Administration, Voluntary Safety Self-Assessment (United States Department of Transportation, 2020). https://www.nhtsa.gov/automated-driving-systems/voluntary-safety-self-assessment. Accessed 1 June 2020.

93 Canada DoT, *Canada's Safety Framework for Automated and Connected Vehicles* (Department of Transport Canada, 2019).

94 B. W. Smith, The public safety case. https://newlypossible.org/files/presentations/2016-07-18_PublicSafetyCase.pdf (2018). Accessed 1 June 2020.

95 General Motors, Safety petition (National Highway Traffic Safety Administration, 2018). https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/gm_petition.pdf. Accessed 1 June 2020.

96 Nuro, Inc., Petition for exemption from certain provisions of Federal Motor Vehicle Safety Standard, No. 500 (National Highway Traffic Safety Administration, 2018). https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/nuro_petition.pdf. Accessed 1 June 2020.

97 C. Luetge, The German ethics code for automated and connected driving. *Philos. Technol.* **30**, 547–558 (2017).

98 B. W. Smith, How reporters can evaluate automated driving announcements. *J. Law Mob.* **2020**, 1–16 (2020).