12-19-2021

Ethics, Safety, and Autonomous Vehicles

William H. Widen

Philip Koopma

Benjamin Kuipers

Marilyn Wolf

Ethics, Safety, and Autonomous Vehicles

Philip Koopman, Carnegie Mellon University

Benjamin Kuipers, University of Michigan

William H. Widen, University of Miami

Marilyn Wolf, University of Nebraska–Lincoln

This roundtable explores the ethical and safety implications of the rapidly evolving technology of autonomous vehicles.

This roundtable discussion reflects a virtual dialog among the authors about ethical issues of autonomous vehicle (AV) design, which each of them has been investigating in different ways. (See "Roundtable Panelists" for more information about the panel.) Philip Koopman is an engineering professor at Carnegie Mellon University who is an expert on AV safety engineering. Benjamin Kuipers is professor of computer science and engineering at the University of Michigan, doing artificial intelligence (AI) research focused on domains of foundational knowledge, including ethics. William H. Widen is a law professor at the University of Miami School

of Law who has been researching the relationship between securities law disclosure and ethics for AV companies related to the decision to deploy AV technology at scale. Marilyn Wolf is Koch Professor of Engineering and Director of the School of Computing at the University of Nebraska–Lincoln; her research interests include embedded computer vision.

COMPUTER: Thanks very much for joining us in this virtual meeting. I think that we have an unusually diverse group to discuss this important topic. AVs have moved from science fiction to advanced prototypes in a remarkably short time. These vehicles introduce new types of questions that the industry has had a limited amount of time to grasp. Hopefully, our conversation today can help to identify some interesting questions as well as avenues for possible answers and further research.

As an opening question, what do we mean by autonomous vehicles or "AVs" for short?

PHILIP KOOPMAN: Let's use the informal definition that an AV is one in which nobody has real-time responsibility for operating the vehicle. If someone inside the vehicle or remotely monitoring the vehicle can be blamed for making a mistake that leads to a crash, it's not autonomous.

COMPUTER: The so-called "Trolley Problem" has become a popular discussion point for the ethics of autonomous vehicle design. Can someone briefly explain the problem for us? How useful is this example?

WILLIAM H. WIDEN: The Trolley Problem is an ethical dilemma where one must make a choice of whether or not to pull a switch to direct a runaway trolley onto a track with one worker and away from a track with five workers, when either choice is fatal to those hit. It is based on scenarios originally presented by Philippa Foot in 1967, though Judith Jarvis Thomson gave it the name Trolley Problem in a 1985 *Yale Law Journal* article.

It is an artificial example of a binary choice with certain outcomes. Most people have the ethical intuition to pull the switch to hit one and not five to reduce loss of life for utilitarian reasons.

KOOPMAN: As a practical matter, the Trolley Problem has distracted attention from much more pressing ethical issues such as governance models for making deployment decisions.

While it is intellectually interesting to consider the Trolley Problem, today's technology is nowhere near the point at which it is relevant to real vehicles that we can build any time soon. It assumes that the vehicle is able to perfectly assess the traffic situation and accurately predict possible

outcomes of actions such as how much damage a low-speed vehicle impact will do to each specific person potentially involved. We're not there yet, and we're not almost there yet either.

WIDEN: The Trolley Problem is a thought experiment for philosophical reflection and not a problem that asks for a real-world answer. Much ink has been spilled in articles that do not understand this.

What most people call the Trolley Problem is really a "trolley case."[1] The original Trolley Problem compared the person at the switch to a doctor deciding whether to harvest the organs of one person to save five. The "problem" was to explain why there is universal

ROUNDTABLE PANELISTS

Philip Koopman is an internationally recognized expert on autonomous vehicle (AV) safety whose work in that area spans 25 years. He is also actively involved with AV policy and standards as well as more general embedded system design and software quality. His pioneering research work includes software robustness testing and runtime monitoring of autonomous systems to identify how they break and how to fix them. He has extensive experience in software safety and software quality across numerous transportation, industrial, and defense application domains, including conventional automotive software and hardware systems. He was the principal technical contributor to the ANSI/UL 4600 standard for autonomous system safety issued in 2020. He is a faculty member of the Carnegie Mellon University electrical and engineering department department, where he teaches software skills for mission-critical systems. In 2018 he was awarded the highly selective IEEE-SSIT Carl Barus Award for outstanding service in the public interest for his work in promoting automotive computer-based system safety. Contact him at koopman@cmu.edu.

Benjamin Kuipers is a professor of computer science and engineering at the University of Michigan. He was previously at the University of Texas at Austin, where he held an endowed professorship and served as the Computer Science Department chair. He received his B.A. from Swarthmore College and his Ph.D. from the Massachusetts Institute of Technology. He is a Fellow of IEEE, the American Association for Artificial Intelligence, and the American Association for the Advancement of Science. His research in artificial intelligence (AI) and robotics has focused on the representation, learning, and use of foundational domains of commonsense knowledge, including knowledge of space, dynamical change, objects, and actions. He is currently investigating ethics as a foundational domain of knowledge for robots and other AIs that may act as members of human society. Contact him at kuipers@umich.edu.

William H. Widen is a professor at the University of Miami School of Law. He received his A.B. in philosophy with honors and distinction from Stanford University in 1980 and a J.D. cum laude from Harvard Law School in 1983, where he was an editor of *Harvard Law Review*. He was a law clerk to the Hon. Levin Campbell on the federal First Circuit Court of Appeals in Boston and then practiced corporate and securities law in New York for 17 years at Cravath, Swaine & Moore, where he was a partner, before moving to the academy in 2001. He is an elected member of the American Law Institute. Contact him at wwiden@law.miami.edu.

condemnation of a decision by a doctor to harvest organs, yet almost everyone believes pulling the switch to sacrifice one to save five in a trolley case is either permitted or mandatory. The challenge is to explain our different ethical intuitions despite a surface similarity that one is sacrificed to save five in both cases.[2]

BENJAMIN KUIPERS: Even though the Trolley Problem assumptions are inappropriate for real-world AVs, the famous "Moral Machine" polling experiment,[3] which has caused so much recent concern, makes exactly those assumptions, a point I put on the table for discussion.

WIDEN: The Moral Machine "experiment" is an exercise in experimental ethics in which millions of people participated in an online poll conducted by folks at the Massachusetts Institute of Technology (MIT) collecting preferences from around the world for choices such as "Would you swerve a vehicle to hit grandma if needed to save junior?" Decisions like this cause concern because it violates the idea that all people should be treated equally regardless of personal characteristics. It smacks of the kind of overprecise utilitarian calculation that sounds good in principle but can rarely, if ever, be carried out in practice. We are concerned not only by the unequal treatment but also by the profound sense that any such attempt is likely to get it wrong in any event. The idea that polling might determine our ethical principles causes concern because we think ethics is not simply a matter of opinion determined by a poll even if we do not think ethical statements are equivalent to factual statements—if we doubt an "ought" may be derived from an "is."

KUIPERS: As for the Trolley Problem itself, in a continuous AV-driving world, perception is uncertain and is best represented as a probability distribution over possible worlds, given the perceptual image provided by the

sensors in the AV. In that (vast) set of possible worlds, the ones that provide only two possible outcomes (kill A versus kill B) constitute a very small subset, with very low probability. Far more likely are other outcomes, including many predicting less harm. Likewise, the outcomes of actions are uncertain, with unintended results being real possibilities. Indeed, Philippa Foot noted back in 1967[4] (when she first started worrying about trolley cases) that the real world is about probabilities and not certainty.

With all this uncertainty, the action that maximizes expected utility (minimizing death and harm) is likely to be one that targets an intermediate outcome, with the largest margin separating the most likely outcome from catastrophe. Even if catastrophe does ensue, the AV tried to avoid it.

However, every young student in driver's training learns a better answer to this problem. When you turn onto a narrow street where a sudden obstacle might be unavoidable, slow down just in case. Driving skill comes not from learning to choose the lesser of two evils but from learning to recognize the "upstream decision points" that avoid the dilemmas entirely.

The Moral Machine experiment defines a "box" around two evils and forces participants to choose who to save based on the situation and personal characteristics of the possible victims. The Trolley Problem does the same thing but considers only the numbers, not personal characteristics of the potential victims. The skilled driver, whether human or AV, thinks outside this box. The responsibility of the AV developer is to make sure that the AV has the necessary knowledge of "upstream decision points" and the skill to act properly as circumstances require.

KOOPMAN: I agree with Ben. What you want is an AV that is "smart" enough to avoid getting into a no-win situation in the first place. The right idea is to anticipate possible danger and avoid it—a classic exercise in defensive driving.

WIDEN: I would not focus on artificially constrained binary choices with certain outcomes. Ben's point about how we teach student drivers captures this essential idea. The constrained binary choice with certain outcomes almost never arises in the real world. As Ben and Phil point out, the real-world problems do not have certain outcomes but are exercises in probability and the technology is nowhere near being able to process this problem in a probabilistic setting. I think we all agree that the Trolley Problem focuses on the wrong things for ethical AV design.

COMPUTER: Any further comments on the MIT Moral Machine survey?

KOOPMAN: As a practical matter, it is unrealistic for an AV to have adequate information to even try such an approach. And it's a machine, so who wants a machine deciding who lives and who dies based on personal characteristics?

WIDEN: I have seen technology at Georgia Tech that can identify pedestrians (at least in a basic fashion) and I have seen the way the system assigns risk to different areas in a dynamic visual scene. I was shown how one could specify any risk weighting that one wanted to use for an individual. This could include assigning different values to different persons in a scene if that information could be provided nonvisually (say, by taking it from a person's cell phone, which we assume each person is carrying). So, the nightmare scenario in Moral Machine does not seem that far off as a matter of technology development.

KOOPMAN: A demo that can do this with fair accuracy might not be far off. One that can do it in real time, at scale, with high accuracy is quite far off. What if hospital staff are assigned a high value in a decision algorithm but only those wearing scrubs are recognized? And then what do you do about all the imposters wearing "safety

scrub" fashion style to improve their odds as pedestrians? I don't see this type of technology as viable in practice. The Moral Machine is unrealistic for AV safety because it assumes that the vehicle has knowledge that is unlikely to be available in a real-world crisis situation, such as the age and profession of a potential victim. (If someone suggests cell phone information be used to support such a scheme, that will create an instant market for spoofing, even if it could be done in real time at scale.)

The German Ethics Commission issued a report in 2017 that expressly takes the defensive driving posture. They prohibit sacrificing noninvolved parties and expressly condemn any use of classification based on personal features such as age and gender, as is done by the Moral Machine.

COMPUTER: Short term, given the current state of machine learning (ML)/AI, what problems do we anticipate?

KOOPMAN: A significant issue with this technology is the unknowns. If you're fundamentally taking an ML approach of training on things you have seen, what happens when you inevitably encounter one of those famous unknown unknowns that you didn't see in training or testing? Worse, what if we find out that the number of unknown unknowns is itself unknowable?

When the AV industry started getting serious funding, I pegged the required safety at 10 to 100 times safer than humans for two reasons. First, every time someone is killed by an AV, the public messaging and litigation arenas are going to see the one victim killed by a vehicle failure (rather than human driver error) but not the people statistically saved. So, it had better be a dramatic reduction.

Second, there will likely be significant uncertainty as to the expected on-road safety during initial deployment. A factor of 10 or more safety margin gives you some room in case the real world turns out to be harder than you thought during testing, which is inevitable.

Where we are now is that we hope that the promise of AVs to save lives will work out. AVs will make different mistakes than human drivers, and nobody yet knows how long it will take to get the balance in favor of AVs. I'd really like to see a credible safety case backed by solid evidence showing that AVs will be at least as safe as human drivers—with ample margin for error—before deploying.

KUIPERS: One perspective on AI and robotics technology, including AVs, treats them just like other potentially impactful technologies such as nuclear power and genetic engineering. We try to think carefully about costs and benefits and what level of understanding we need before deploying the technology. We all expect problems with assessing the true safety of AVs as the time approaches to decide whether to deploy AVs at scale. Developing standards and metrics will be a central problem for the AV industry.

There is another perspective on AI (including robotics and AVs). We are creating agents that perceive the world, make their own models of the world based on these perceptions, and make their own ethical decisions about how to act—what move to make next. This requires AI system developers to understand what ethics are and how ethical knowledge can be represented and used by a robot. This is more complex than merely considering the consequences on a cost/benefit basis of whether a deployment of an impactful technology results in a positive balance of utilities. This utilitarian calculation "simply" requires ethical thinking by human developers. I would like an explanation of how the robot represents and uses ethical knowledge before deployment.

WIDEN: I am worried about what disclosure is appropriate for purchasers of an AV. If part of the functionality of the AV is "rule based" as reflected in an algorithm, I think I have a better idea how I might accurately describe that aspect

of the AV as disclosure for consumers. To Ben's important point about deployment, I can add that the AV industry already shows signs of an inability or unwillingness to expressly identify standards for deployment. This problem appears in the recent U.S. Securities and Exchange Commission (SEC) filing for Aurora Innovation, Inc. (a problem that I have recently addressed in an essay about SEC disclosure).

KOOPMAN: Using ML techniques breaks the traditional design "V" process (top-down design on the left side of the V, bottom-up validation on the right side of the V) that has been the foundation of computer-based system safety engineering for more than 20 years. While you can try to make an ML–based process look like a V, you can easily weaken or lose explicit design intent (which is now implicit in what is learned). So, it's difficult to know if you've done the right testing on the right side of the V to make sure that design intent has been achieved. In practice, it is common to see surprises in deployment that most testers would never have dreamed were relevant to safety-critical behaviors.

MARILYN WOLF: Methodologies like ISO 26262 (an automotive functional safety standard based on the V process) wrap more general assurance methods around analytic methods like control theory that characterize specific cases such as step response. These methodologies rely on characteristics like continuity allow inference of system behavior in parts of the design space that haven't been directly analyzed. Modern ML systems don't have those characteristics—a very small change in input can result in a completely different output.

KOOPMAN: As an example of a surprise, my team was working with a commonly used computer vision system and discovered weaknesses with identifying people wearing high-visibility clothing (yellow raincoats and

high-visibility yellow/lime vests). These were not just random dropouts. Rather, there were long video sequences with a little bit of random visual noise in which yellow-clad people were obviously there and taking up a lot of the image field to human eyes but were simply not detected by the vision system. In other words, high-visibility clothing was essentially camouflage for a vision system.

While you can fix problems after you notice them, noticing them is the hard part because they can be really unexpected. It doesn't matter what the designer thinks might be an edge case. What matters is characteristics that humans might not think twice about that break the ML. This is especially problematic for smaller populations of highly vulnerable road users such as wheelchair riders who don't appear often in randomly sampled data.

A related insight is that an ML–based system does not tell you if an object is actually a person. It tells you if an object is statistically similar to all the people it has seen before. If you are a person who doesn't look ordinary, you are at risk of not being seen. "Ordinary" includes not only biases we as humans typically consider (skin color, size, use of mobility aids) but also more subtle contextual information such as clothing color and standing in front of a background with strong vertical edges.

COMPUTER: When is an autonomous system acceptable to deploy?

WIDEN: The recent registration statement filed with the SEC for Aurora Innovation's going public transaction[5] reveals a fundamental problem: the AV industry does not want to tell you what its standard for deployment actually is: will AV companies deploy without a safety driver only if there is great confidence that a machine driver is safer than a human driver, or will they deploy when machine drivers are less safe but on the hope that, in the future, autonomous vehicles will be more safe? Deployment on hope is a utilitarian

"losses now, gains later" justification that assumes that engineers can "patch their way to perfection"—a world in which there are zero traffic fatalities. It looks a lot like using people on the highways as guinea pigs.

KOOPMAN: Vehicles shouldn't be deployed until they meet industry safety standards and have a credible explanation for why they will be acceptably safe. Right now, that means at least ISO 26262, ISO 21448, ANSI/UL 4600, and an applicable cybersecurity standard as well. These are industry-created, flexible, consensus safety standards issued by accredited standards development organizations. However, many AV developers won't commit to following them in any substantive way, and at least one developer is actively pushing back against them claiming they are instead following some unspecified better way. As with aircraft, safety should be a given (that is, the same for all industry players) and competition among AV companies should hinge on other factors.[6] Is it too much to ask the AV industry to follow its own safety standards?

COMPUTER: What approaches to enhancing safety may be appropriate for the short to medium term?

KOOPMAN: Short term, driver assistance systems such as automatic emergency braking are likely to provide much more net safety than AVs because they can be deployed on all vehicles right now. These features are also improving human driver safety, so AVs are chasing a steadily improving human driver safety target.

A bigger issue from an ethical point of view is that risk management (from an insurance or corporate profit point of view) and safety assessment are cousins but not the same thing. Companies are financially incentivized to manage and reduce financial risk. But there are situations in which maximizing profit via risk management does not result in what many would consider "acceptable" safety. Those

situations tend to occur when the cost of a design choice is higher than the expected cost of settling litigation for an expectation of a small number of deaths or severe injuries.

The Pinto gas tank fire cases were a poster child, but this has happened in more than that one case. The tricky part tends to be that companies often underestimate how often a bad thing can happen, and don't fully account for potential reputational damage from even a few high-profile loss events. If the industry gets this wrong in a major way, they risk losing their traditional self-certification privilege.

KUIPERS: I think your "bigger issue" is the heart of the problem. Engineers and corporate managers are taught methods for utility maximization, but they are not taught how to be sufficiently careful and thoughtful about how they define utility. When taught, utility is defined in straightforward ways, like dollar costs. One might get sophisticated and use expected discounted dollar costs. But reputational costs, or more generally the value of trust, is difficult to fit into this maximization framework, so it is tempting to leave it out. That approach even gets valorized with slogans like "Greed is good."

WIDEN: Ben is right. Greed may be valorized when the interests of society are aligned with the interests of an AV business, but the case of AV technology is not the same as an Adam Smith world with the goal of producing goods and services at the right prices in the right amounts. A poorly designed AV creates risks for the public in ways that a generic mis-priced good or service does not.

WOLF: The Tempe accident is an interesting example for our decisions on when a technology is deployable. The National Transportation Safety Board (NTSB) report[7] says that the original Volvo ADAS system was tested under simulation; the results showed that the Volvo system would have avoided collision in 17 out of 20 scenarios and

reduced impact speed to fewer than 10 mph in the other cases. Unfortunately, the Volvo ADAS systems were disabled when the Uber automated driving system was operational.

KOOPMAN: It's important to realize that for the Tempe Uber ATG fatality technology was involved, but a deeper root cause was a poor safety culture and no real safety management system (SMS). To paraphrase the NTSB chair's opening remarks at the hearing: you don't have to wait until it's your company that has a fatal crash to create an SMS.

COMPUTER: What rules should the AV follow while driving?

WIDEN: Looking at AV technology as only an accident reduction or life-saving scheme is incomplete. We need to understand that AVs will operate within an "institution," which, in this case, is framed by highway rules and regulations. The first goal of an AV should be to follow the rules of the road. This tempers all the crazy scenarios about trolley problems and ethical dilemmas because any actions need to be evaluated against this background institution. We may need set of rules for AVs like Asimov's rules for robots:

1. Subject to rule 2), obey all traffic laws (the institutional rules condition).
2. Violate rule 1) only if it is necessary to avoid an accident and the rule violation does not place a human at risk of harm (the lane changing condition).
3. Subject to rules 1) and 2), operate to reduce accidents/collisions to the maximum extent possible (the braking/distance maintenance condition).

Once you understand that the rules of the road create the institutional system in which the AV must operate, the prioritization of pure utility maximization in individual cases makes less

sense. Utility maximization can take place, but only it can be done within the structure of the individual rules. It also does not require rules that might create bad incentives, like "hit the motorcyclist without the helmet" or "sometimes drive on a sidewalk." Pedestrians should feel safe walking on sidewalks. The March 2021 Automated Vehicle Safety Consortium Best Practice for Metrics and Methods for Assessing Safety Performance of Automated Driving Systems does mention following the rules of the road as part of the goal.

WOLF: Keep in mind that Asimov created his laws as framing devices for his story. His purpose as a writer was to explore the ambiguity inherent in seemingly reasonable robot laws.

KOOPMAN: Simple rules like that are nice in principle but get messy in practice. Real-world traffic laws are treated as more of a guideline than absolute rules. There is a lot of room for driver discretion. Do you put two wheels over the center line on a two-lane road to avoid running over a downed power line? Do you do it to give someone changing a flat tire more room if the road is otherwise empty? For a rules-based approach to work, you're going to need rules of the road with a lot less reliance on "do the right thing" and "drive friendly."

COMPUTER: How do corporations in this space approach these issues?

WIDEN: Aurora's SEC filing states that it will "operate with integrity," and that "we do the right thing." Their stated goal is to build "trust." At the same time, the Form S-4 says Aurora will "[b]e reasonable" but the scope of good judgment is limited to "always have the best interest of the company and our partners in mind." Of course, that is just the traditional corporate fiduciary standard. Some view the corporate fiduciary standard as including a requirement to break laws if that is

the economically rational thing to do because it increases shareholder value.

The mistake that I see reflected in the S-4 is the faulty inference that accident rates will go down because one has eliminated the human errors. You need to know the frequency of machine actor errors that enter the system. That, it seems to me, is what the AV companies do not have a reasonable basis to claim. Yet, they have an "education campaign" to convince consumers of the benefits of AV technology when the benefits are merely a hope at this stage, and not a reality. The AV companies need to convince the public that AVs are safe when they cannot demonstrate safety—but they need this perception to begin deployment, or the public will revolt.

COMPUTER: Given fundamental advances in ML/AI, what ethical guiding principles are appropriate for the longer term?

KUIPERS: One of the classic questions on a driver's ed test is: What do you do if you see a ball rolling into the street, with no people visible anywhere? The obvious answer is to stop (or at least slow down drastically) because there could well be a child chasing that ball. AV technology needs to achieve this degree of sophistication before the public might really have confidence in the technology. The public needs to "trust" that AVs are truly safer than human drivers. How can you trust an AV that could not pass a driver's ed test? Putting an AV on the road, which cannot pass a driving test would appear to violate basic ethical principles.

KOOPMAN: We need to decide if we want explicit ethical mechanisms designed into a machine.

With a pure end-to-end ML–based approach any ethical values are implicit in the training data because the AV is taught by example. Many developers are proposing separated safety monitoring systems that for example estimate safe following distance based on

Newtonian physics. We might see some ethical guidelines embedded in those boxes, at least at the level of whether it is better to crash the vehicle into a utility pole and rely on internal safety devices to protect passengers rather than strike a vulnerable road user.

Fortunately, we have quite a bit of time to work with for the long term while we try to reduce the number of fatalities and severe injuries that don't require complex ethical decision making.

KUIPERS: I propose that the relevant attributes, when deciding what to do in a situation, are not utility, but trust (and trustworthiness). "Trust is a psychological state comprising the intention to accept vulnerability based on positive expectations of the intentions or behavior of another."[8] We should evaluate robot or AI behavior in terms of whether the agent deserves our trust and demonstrates trustworthiness.[9]

As a pedestrian, I am *not* prepared to accept the vulnerability that an AV could decide to deliberately kill me because of its internal evaluation of the other potential victims in its path. Rather than accept that vulnerability, I would join forces with other like-minded pedestrians (and there would be plenty) to ban this innovation entirely, regardless of its potential benefits.

I would want to trust that the AV has been implemented with superhuman abilities for defensive driving: identifying the upstream decision points that will avoid deadly dilemmas. In the (therefore extremely rare) case of being faced by a choice among evils, I would want to trust that the system has a superhuman ability to minimize human fatalities and injuries, taking into account perceptual and action uncertainty.

WIDEN: I think the "trust" point presents a good point of entry to consider the following. Why do we trust other people? We trust other people because we believe that, fundamentally, they are "just like us"—more so for family and relatives. We believe that a normal human being will respond to basic situations the same way that we would respond. This is because we have a certain sympathy or empathy for other persons (even though we can never share their subjective experiences). The other humans are thus not aliens or "other." This is one basis for developing a relationship of trust.

If this is right, it might explain why we have trouble trusting a machine actor. We do not have the same empathy, or sympathy, or feeling that it will behave as we would behave on our best behavior.

KOOPMAN: And we also know that the machine actor does not have empathy or sympathy for us.

WIDEN: Some of the trust in strangers might be based on an expected rational response to deterrence by a human actor. That is how Oliver Wendell Holmes advised one ought to understand the law. Do not look at law as morality. Look at law from the viewpoint of the "bad man" who only cares what is legal and not moral. So, the law includes disincentives for bad actions. I am not sure this idea of deterrence applies to AVs.

KOOPMAN: On the other hand, fear of consequences is baked in pretty well to most human drivers, even if they are prone to bending the rules. If an AV has no fear of consequences, we need to find a basis for trust other than good intentions (does an AV even have that type of intention?).

WOLF: As for trust, people will always rely on it. But it seems to me that we, as professionals, have an obligation to provide some complementary analysis based on scientific and engineering principles.

KUIPERS: The institutional rules need to be well designed, but the critical point about the trust perspective is how well individual decision makers trust that other people will follow those rules. For example, there is a rule against driving through a red light, but in many jurisdictions, people often drive through "orange" lights (as yellow turns to red) or even later. This teaches everyone that you cannot trust a green light to allow you to drive forward and makes driving less safe and efficient. (The European red-yellow-green transition helps correct this problem.)

Recall that trust is willingness to accept vulnerability, confidently expecting not to be exploited. This pays off in two related ways:

1. Cooperation with a trusted partner pays off much better than individual efforts.
2. Being able to count on social norms (and institutional rules) being followed provides valuable assumptions when planning one's activities, saving on defense and failure recovery.

Trust might act as a replacement for certainty. Trust expresses confidence that a person or institution will behave in an expected way. If the trust is justified and the social norms and rules are followed, more efficient social results are possible. If the public trusts the AV companies to "do the right thing," they will trust the AV's programming and perception of the environment, not making a fuss over deployment in spite of uncertainty about the future.[10] But the trust needs to be justified.

WIDEN: This focus on trust is exactly right. I see two issues of trust. Do we trust what the AV will do? Do we trust the AV makers that the AV will perform as advertised? To assure people about my future conduct so that I am trusted, it may not be enough merely to state my principles. People need to see that I actually follow those principles (an idea I get from Robert Nozick). But it is hard for ordinary people to observe how an AV company is following safety principles in development of complex technologies like AVs. It is even harder if I use very vague standards like

"sufficiently safe" rather than "much safer than a human driver."

WOLF: Trust is an emotional state that is deeply rooted in human behavior, perhaps stemming from the mother–child relationship. Trust decisions are often made without regard to rationality—both in when and where to bestow trust and when to remove that trust.

KOOPMAN: One factor of trust is the reputation of the company. Preferably based on a track record and not simply saying "we're really smart, so trust us." Another is word of mouth from personal and social circle experiences.

A third, more technically substantive basis for trust is conformance to standards with attestations from independent parties (think Good Housekeeping Seal or TÜV testing). However, the auto industry is essentially unique in that they "self-certify" for safety and do not have a strong historical track record of following (or at least publicly stating that they follow) their own industry computer-related safety standards.

Then there is regulation. Historically, aviation regulators have been proactively involved in safety decisions for aircraft while they're being designed. But for software-intensive functions in cars the U.S. National Highway Traffic Safety Administration (NHTSA) has relied primarily on recalls and other reactive measures. Once there is no human driver to blame for crashes, NHTSA is going to have to pivot hard into dealing with software safety up front.

What other trust factors am I missing?

WOLF: How about self-promotion?

KOOPMAN: A significant problem with trust is that it can initially be earned too quickly. An hour ride in a seemingly safe vehicle can lull people into complacency—even though life-critical dependability requires tens of millions of hours of exposure to establish statistically significant

life-critical dependability. Then there can be a backlash once trust is broken.

KUIPERS: A suggestion for building trust in AVs: detect and avoid deer collisions. Collisions with deer are particularly hard to anticipate and avoid. From the human driver's point of view, a deer simply appears in the path of the car with no warning at all.

An AV has cameras, lidar, and radar that sense 360° around the car, and the system can watch carefully *all* the time. A radar mounted under the car could look under parked cars and detect deer moving toward the road that would be literally invisible to a human. Demonstrating that an AV can successfully avoid deer collisions would encourage trust that the AV could also avoid collisions with people.

WOLF: Machines don't suffer from some of the distractions prone to people. However, they operate with finite computing resources that mean they may end up missing some events. Given the latency constraints required for driving decisions, off-loading decisions to the cloud is impractical. The AV computing system may need to prioritize some tasks over others, taking into account factors such as importance and timing constraints.

COMPUTER: What professional responsibilities do we have as system designers?

KOOPMAN: Do we trust companies at least partially motivated by potential windfall profits or IPO/SPAC (an "IPO" is an initial public offering and a "SPAC" is a special purpose acquisition company that is used to take a company public) valuations to decide for themselves what risks are acceptable when publicly testing such technology? Is potential legal liability sufficient incentive for them to act in a safe, responsible manner?

WIDEN: If you are asking about system designers at an AV company, they need to recognize that there is a moral

hazard—their decision making about advertising and deployment may be clouded by financial exigency—this is a worry I have about Aurora if no corporate procedures are put in place to protect the integrity of the decision process. I suspect there is only so much a line engineer can do in the form of whistleblowing.

WOLF: Shouldn't someone be concerned that we don't have a safety methodology for safety-critical autonomous systems?

The Federal Aviation Administration put out a recommendation for the use of ML in aircraft. They propose wrapping the ML component in a control loop that acts as a limiter. This would provide some benefits for ML but fundamentally limits the scope of those benefits. And I am not convinced that this approach eliminates safety problems. Consider, for example, an ML system that decides to bang the joystick randomly around the four corners. The control system may not be able to compensate.

KOOPMAN: In traditional safety-critical systems, the idealized framework for safe design is that the designers fully specify the system accounting for all possible situations. This is adapted for use in the messy real world by delegating to a human the job of risk mitigation beyond the scope of what the system can handle. Easy ethical issues are designed in intentionally, and hard ethical choices are often kicked up to a human operator.

WOLF: We may be confusing the technologies AI/ML and the application autonomy in some situations, particularly when we complain about limitations. The definition of autonomy is important because it tells us when we have to apply some new engineering methods.

We seem to think that traditional safety-critical system design is insufficient for AVs. Is that due to the underlying AI/ML technology? Or is it because we expect the vehicle to perform in a wider range of situations?

Engineering traditionally analyzes a machine's response to a well-understood set of inputs or forces; we then generalize to use cases. Do we need to do the same thing for AVs by restricting where they operate autonomously, perhaps never getting to Level 5? (General Motor's driver assistance system limits itself to certain premapped highways.) Whatever new engineering methods we decide we that we need, how do we train students to understand when they have entered that new territory and need to apply an additional level of methodology?

KOOPMAN: Without a human operator we can put a safety envelope around some potentially dangerous AI/ML actions to help mitigate risk. However, we're still not sure how to do this effectively for some functions such as object classification. (Pedestrian or tree? We've seen bare legs and brown pants on people mistaken for tree trunks.) System designers have a responsibility to think very carefully about ensuring that the parts of the AI/ML system that don't have safety envelopes are fit for purpose. If they are relying on big data arguments, this extends to making sure that the data collection and curation processes are similarly robust enough to trust them with people's lives.

WOLF: I keep coming back to the use case issue. Do we think that we know how to create an AV that we would want to have driving through our neighborhood school zone when school lets out? Or would we be better off—at least for the time being—sticking with the still-challenging but relatively simpler highway driving case? There is an abstract question of how we create autonomies, and so on. But people are designing these cars right now.

COMPUTER: What should we teach our students about ethics, safety, and AVs?

WOLF: We need to think more about what we teach our students. It seems to me that teaching students even a little bit about ethics should change their thinking. First, that there are principles to help guide our decisions, such as utilitarianism versus giving everyone an equal chance at survival. Second, that different principles may lead to very different outcomes: hitting one person not five versus somehow giving everyone a chance at survival by flipping a coin as suggested by the philosopher John Taurek many years ago.[11]

WIDEN: The MIT Moral Machine experiment is just a newish discipline called *empirical philosophy*. It has its place as a point of information, mostly about variance across cultures. No serious person thinks you get ethics from a poll. I would start with that insight. And then focus students on the incredibly difficult task of actually performing utilitarian calculations. Despite the difficulty of these calculations, society often needs to justify a decision using a cost/benefit analysis. But on the other side, there are certain personal rights, which are sacrosanct and may not be overcome by a utilitarian calculus—you can't harvest organs, for example.

The difference between philosophical consideration of an ethical dilemma and deployment of an AV, is that in the thought experiment only feelings get bruised if the experiment sends an interlocutor into a state of aporia (a fancy Greek term for confusion). On the highway, somebody gets killed. Theory meets practice in the road test.

KOOPMAN: Teaching ethical principles is important, but we also need to make sure our students are equipped to deal with the system-wide implications of safety.[12] Right now, we have a situation in which the Silicon Valley ("move fast and break things") culture is trying to work with the automotive culture ("probably it was driver error"), the AI/ML culture ("99% is amazing"), and the computer-based system safety community ("99% is indeed impressive for that technology. But life-critical is more like 99.999999% per mile"). Students need to be able to reconcile all these different approaches to dependability in their heads at the same time.

KUIPERS: Cooperation delivers better rewards than noncooperative effort, but it depends on trust among the partners. The "Prisoner's Dilemma" illustrates this: when each individual tries to maximize reward and eliminate vulnerability, the outcomes are bad for both the individuals and the group. A good outcome requires justified trust. Social norms like "drive on the right" are a kind of generalized cooperation across the whole society. When we can trust them, they make everyone's travel safer and more efficient. The bottom line is trust, earned through trustworthiness.[13] Both engineers and management of AV companies need to focus on structures that promote trust in just the right way.

COMPUTER: Thank you, all, for this excellent discussion. Autonomous vehicle development is moving very rapidly; it also refers to some fundamental concepts in ethics, law, and AI. ▣

REFERENCES

1. R. Lawlor, *The Ethics of Automated Vehicles: Why Self-driving Cars Should not Swerve in Dilemma Cases*. Res Publica, July 6, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s11158-021-09519-y

2. W. H. Widen, "Autonomous vehicles, moral hazards & the 'AV problem.'" School of Law, Univ. of Miami Working Paper Series. [Online]. Available: https://ssrn.com/abstract=

3. E. Awad et al., "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018. doi: 10.1038/s41586-018-0637-6.

4. P. Foot, "The problem of abortion and the doctrine of double effect," *Oxford Rev.*, vol. 5, pp. 5–15, 1967. (Reprinted in P. Foot, *Virtues and Vices: And Other Essays in Moral Philosophy*, 2002.) doi: 10.1093/0199252866.003.0002.

5. "Reinvent Technology Partners Y. FORM S-4 registration statement," Securities and Exchange Commission, July 15, 2021. [Online]. Available:

https://www.sec.gov/Archives/edgar/data/1828108/000119312521216134/d184562ds4.htm

6. "Credible autonomy safety argumentation," in *Proc. Twenty-Seventh Safety-Critical Syst. Symp. (SSS'19)*, York U.K.: Safety-Critical Systems Club, 2019. [Online]. Available: https://users.ece.cmu.edu/~koopman/pubs/Koopman19_SSS_CredibleSafetyArgumentation.pdf

7. "Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona," National Traffic Safety Board, Washington, D.C., Accident Rep., NTSB/HAR-19/03, PB2019-101402, Mar. 18, 2018.

8. D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy Manage. Rev.*, vol. 23, no. 3, pp. 393–404, 1998. doi: 10.5465/amr.1998.926617.

9. B. Kuipers, "How can we trust a robot?" *Commun. ACM*, vol. 61, no. 3, pp. 86–95, 2018. doi: 10.1145/3173087.

10. S. Rose-Ackerman, "Trust, honesty and corruption: Reflection on the state-building process," *European J. Sociol./Archives Européennes De Sociologie/Europäisches Archiv Für Soziologie*, vol. 42, no. 3, pp. 526–570, 2001. doi: 10.1017/S0003975601001084.

11. J. Taurek, "Should the numbers count?" *Phil. Pub. Aff.*, vol. 6, no. 4, pp. 293–316, Summer 1977.

12. P. Koopman, "AV Safety" (Embedded System Lecture Notes and Presentations Series). [Online]. Available: https://users.ece.cmu.edu/~koopman/lectures/index.html#av

13. B. Kuipers, "Perspectives on ethics of AI: Computer science," in *Oxford Handbook of Ethics of AI*, M. Dubber, F. Pasquale, and S. Das, Eds. Oxford, U.K.: Oxford Univ. Press, 2020, pp. 421–441.

MARILYN WOLF is the Elmer E. Koch Professor of Engineering and director of the School of Computing at the University of Nebraska–Lincoln, Lincoln, Nebraska, 68588, USA. Contact her at mwolf@unl.edu.