

Safety Ethics for Design & Test of Automated Driving Features

Philip Koopman Carnegie Mellon University
William Widen University of Miami School of Law

Abstract— The highly tool-intensive design and validation of automated driving features brings with it significant opportunities to support ethical practices related to safety for testing and lifecycle support. This article shows how some basic principles from the IEEE 7000 standard on ethical concerns during system design can apply.

Keywords: automated driving, safety, ethics, IEEE 7000

■ **The promise** of automated vehicle technology is to eventually improve safety by eliminating human error from driving in the long term, and helping drivers avoid collisions in the near term. However, a long road remains to achieving the goal of a scalable fleet of completely autonomous vehicles. Meanwhile, in the rush to develop the technology, significant ethical considerations are being left by the wayside.

At first blush, it might seem that ethical issues largely involve setting public policies as to how safe might be safe enough to deploy, and deciding what constraints (if any) to place on machine behavior that might inappropriately favor occupants vs. pedestrians in impending crash situations.

However, other concrete ethical issues are playing out that directly concern design, test, and lifecycle support practices. Without direct engineering support, significant ethical issues regarding safety will go unaddressed. We use three examples to illustrate how the principles in the IEEE 7000 standard for ethical design practices apply to designers and tool vendors working on autonomous vehicles.

ETHICAL VALUE REQUIREMENTS

The IEEE 7000 Standard on ethical concerns during system design establishes a set of processes for considering ethical values as part of system design.¹ A key aspect of that standard is to outline an ethical values elicitation process. That process creates Ethical Value Requirements (EVRs) and Value-Based System Requirements (VBSRs).

Per IEEE 7000, an EVR is an organizational or technical requirement related to one of many stakeholder viewpoints that must be considered in the ethical analysis process. The variety of stakeholders for a socio-technical system such as an automated vehicle operating on public roads is wide, ranging from the manufacturer, to automated vehicle riders, to regulators, to other road users, to users of other transportation modes that might suffer economic impact from diversion of mass transportation revenue.

EVRs are likely to be product-level requirements framed per ethical attributes, such as utilitarian (benefits and harms), virtue (character traits and beliefs), and duty (obligations to oneself and others). An EVR might be a requirement to drive courteously, avoid imposing substantive additional risk on

vulnerable populations, or be at least as safe as an unimpaired human driver under comparable conditions.

To translate into design outcomes, EVRs must be mapped to VBSRs. VBSRs are more technical requirements related to typical design and test practices. For example, a set of behavioral rules needs to be specified to make a “drive courteously” EVR actionable as an engineering task. A corresponding VBSR would be associated with an objective validation strategy.

Typical system engineering requirements for automated driving features might be related to whether the vehicle can make progress driving in a complex environment, and whether it follows traffic rules. VBSRs are those system-level requirements that link to EVRs. For example, a VBSR might be a specific very infrequent target for fatal crashes, which would be one component fulfilling an EVR of a societally acceptable definition of “safe enough” for operation on public roads, regardless of whether some or even all cars involved in loss events are driven by computers. (The topics of “safe enough” and responsibility for safe driving are important, but too complex to consider fully herein.²)

VBSRs capture less obvious ethical considerations that might not be captured in an engineering process that does not take advantage of the insights in IEEE 7000 to perform a broad ethical requirements analysis. Some ethical values listed in IEEE 7000 Annex G that are applicable to vehicle automation features include personal autonomy (e.g., independence and mobility), control (e.g., justifiable rule-breaking), fairness (e.g., lack of bias in risks and benefits), sustainability (e.g., promoting a long life for resource-intensive manufactured equipment), transparency (e.g., openness regarding safety tradeoffs in light of company responsibility to shareholders to turn a profit), and trust (whether the manufacturer will accept accountability when called upon to do so).

In today's world, ethical values related to fairness must specifically include consideration of social justice principles and concerns.³ An additional topic of ethical concern is data privacy, but we do not treat that in detail because it is not unique to automated vehicles.⁴

We use three examples to illustrate how EVRs can – and should – influence the engineering design and test process: public road testing, lifecycle support, and the deployment governance process.

ETHICAL PUBLIC ROAD TESTING

An obvious ethical concern for public road testing of immature automated vehicle features is the risk of harm to other road users. This might be mitigated by conformance to the automotive industry's SAE J3018 road testing safety standard that provides guidance for ensuring effective safety drivers supervise such testing.⁵ (The sad state of autonomous vehicle ethics is such that, as of this writing, no company currently doing public road testing has claimed conformance to J3018.) Even if companies were to conform to their own industry standard for safe road testing, additional EVRs should be addressed to issues such as fairness, social justice, and transparency.

One example of a fairness EVR for testing is motivated by the ethical principle that it is wrong to allocate costs disproportionately to at-risk communities, even if they will also benefit from a technology that is expected to eventually benefit society in general. This EVR should trace to VBSRs for the road testing plan. As an example, testing in low-income or historically disadvantaged parts of a Metropolitan Statistical Area (MSA) might be attractive because of challenging road features. There might also be a financial incentive to test there because the expected cost of compensating for an accidental loss of life (or reduction in earning power due to injury) is lower in low-income areas in the event of a testing mishap. A VBSR might limit time spent in such areas to a small fraction of total testing time, even if that led to testing campaigns that were not as efficient or had a higher expected net cost.

Testing in other areas at particular risk might be limited, including school zones, playground areas, special event sites, and areas near institutions for disability-related mobility skill training. Other considerations might include minimizing use of key emergency response vehicle routes to avoid having a disabled or confused autonomous vehicle block those critical routes.

Designers and toolchain makers can support testing fairness VBSRs by incorporating test planning features that are responsive to equitable concerns about which locations are being exposed to testing risk. For example, the US Government has an online map that identifies areas of persistent poverty and historically disadvantaged communities.⁶ A VBSR might combine this information with local government-supplied information to designate select areas for special risk reduction.

An additional dimension to minimizing the risk of unavoidable testing in designated areas might be to ensure an especially high performance of safety drivers in designated areas (e.g., use a more expensive arrangement of two safety drivers in a vehicle instead of just one; testing only with fresh crews at start of shift), prohibit uncrewed testing in designated areas, and test at less risky times (e.g., minimize testing right after school lets out). Maximizing use of simulation to displace road testing that is not absolutely necessary can also help, but should be used for all road testing, and not just road testing in designated areas.

Uncrewed testing provides little, if any, additional scientific information regarding safety beyond that obtainable with crewed testing, so long as vehicles track whether safety drivers needed to use the controls during a test cycle. Nonetheless, firms have an investor and public image incentive to show they can operate crewless on public roads. VBSRs should at the very least be in place to require only crewed testing in highly sensitive designated areas.

From an optimization point of view, VBSRs add additional constraints for test route planning. The objective for testing might be to collect a particular number of certain types of scenarios on any particular road testing session. In the absence of such VBSRs, financial incentives and time-to-market pressure might well result in a disproportionate risk being imposed upon already vulnerable populations. A VBSR regarding time spent in designated areas adds a test optimization constraint of reducing or eliminating certain types of testing in those areas.

ETHICAL LIFECYCLE SUPPORT

Automated vehicle features will need continuing support for the life of the vehicle. The more advanced the feature, the greater the need for support. New objects and events will become part of the vehicle's operational design domain, traffic rules will change, high definition maps will need to be updated, and new weaknesses will be discovered well after initial deployment.

As the average age of operational road vehicles continues to extend past its current 12 years (with many vehicles at 20 years old or more),⁷ this will fundamentally change the support strategy and costs for automotive manufacturers and supply chains. New architectural approaches for automotive electronics will be required to support dependability and potentially a more modular, upgradeable approach to electronics to ensure long-term maintainability (e.g., Kopetz 2023⁸).

EVRs that address sustainability, trust, personal autonomy, and fairness will be important. Particularly important will be the effects of late-lifecycle support for older vehicles typically purchased by low-income owners on the used car market, with many older vehicles exported for use in third-world countries. There is a real risk that the safety features promised by the manufacturers will only be available to the rich owners of new cars in the first few years of their lifecycle.

It is unclear what the incentives might be for a car company to actively support automation features over a 20+ year vehicle lifecycle when their sales and factory service center maintenance cash flow is concentrated on the first few years of ownership. Example issues include:

- Will safety-critical software updates be available for the full viable life of the equipment? If a manufacturer limits the supported operational life of its vehicles, low-income vehicle owners might not have access to safety features with discontinued support, or might be left with vehicles having unsafe or inoperable automated driving capabilities.
- How can consumers trust repair operations performed by sources other than manufacturer shops? Alternately, will manufacturers have a monopoly on repairs and critical repair materials that makes repairs unaffordable for low-income owners of older vehicles? Will cars disable themselves if not regularly maintained by the manufacturer in the name of ensuring safety?
- Will subscription costs to update data such as mandatory high-definition map data feeds be so expensive that low-income vehicle owners will not be able to afford to turn on safety features that would otherwise be available in a vehicle they purchased used? In an era of monthly subscription fees for heated car seats, such questions need to be asked.
- What happens when an automated vehicle manufacturer goes bankrupt or simply decides to terminate support for older vehicles? Do the cars stop running? Do they lose key safety features? Do their dashboards simply go dark?

Effects on low-income car owners can go beyond whether their vehicle has the latest automated capabilities and safety features activated. For example, if the manufacturer disables a used car it no longer wishes to support, the owner still must pay off the outstanding loan balance used to purchase it.

Benefits touted by the autonomous vehicle industry include each rider's own personal autonomy via being able to operate a car without needing to be qualified for a driver license. Given that the target population for such promises tends to include those who are retired and live in rural areas with poor transit access, that promise will go largely unfulfilled if self-driving features are only available to the rich.

Addressing VBSRs for lifecycle support will require careful consideration of continued availability of automated vehicle capabilities as well as safety. At the very least, the vehicle should be safely operable even if its stream of data updates is turned off for some reason.

There will also need to be a mechanism to ensure sufficient information is available to enable third-party maintenance and support if manufacturers are unable to or choose not to provide it themselves. (These issues might be mitigated by adopting a fleet ownership model in which individuals do not actually own specific vehicles. However, manufacturers are already deploying limited automated driving systems in individually owned vehicles, so this issue will need to be addressed one way or another.)

There are broader social policy implications for whether it is acceptable for used vehicles to have some of their safety features disabled, but those go beyond specific technical activities and into corporate policy issues. Should an industry standard or a government regulation mandate a minimum useful life for driving automation features that extends far beyond typical warranty lengths? This might become necessary if an EVR suggests that purchasers of used vehicles should retain access to data-hungry vehicle automation features for the full vehicle lifecycle.

ETHICAL DEPLOYMENT GOVERNANCE

A third example area of ethical consideration is governance of the deployment decision: who decides when it is time to release a new version of vehicle automation features onto public roads, based on what criteria? This raises issues of transparency and potentially other issues – but without transparency there is no way of even knowing what other ethical issues might exist.

A significant challenge is being able to communicate to non-technical stakeholders whether automated vehicle features are acceptably safe, as well as what the basis for the belief in safety might be. The current situation in the industry largely consists of promises of eventual safety paired with fiercely defended opacity regarding any data that might

actually support or refute the veracity of those promises.

A crucial EVR for ethical deployment of automated vehicle technology is being able to explain the approach to ensuring and validating safety. This should lead to VBSRs such as:

- Disclosure of the precise definition of acceptable safety being used in the release process, including objective goals for defined metrics.
- Creation of a safety case, a high-level portion of which can be made public to explain to stakeholders why the automation features are believed to be safe and based on what data. Independent confirmation of conformance to the ANSI/UL 4600 standard can be offered as additional assurance that undisclosed portions of the safety case are acceptable.⁹
- Disclosure of the decision process including metrics used for a deployment decision.
- An archival record of quantified metrics that were used for each such deployment decision. (Safety Performance Indicators required by UL 4600 are suitable for this purpose.)
- Recording safety incidents in a cumulative hazard log that also includes root cause analysis and traceability to corrective actions.
- Ensuring that data collection, retention, and analysis consider privacy concerns.

Vehicle companies currently make technology deployment decisions with no consultation with external stakeholders beyond meeting imposed insurance and administrative requirements. Ethical deployment governance requires significantly more consultation with and transparency to a wide variety of additional stakeholders.

A key EVR question is how heavily speculative future benefits might weigh in deployment decisions. A utilitarian case for deployment of long-haul automated trucks might be made based on the projected shortage of professional truck drivers – even though the automated trucks might potentially be more dangerous than human truck drivers when initially deployed. One might believe that it is unethical to make utilitarian calculations on a gamble that technology more dangerous than human drivers in the short term will, perhaps, eventually be safer than human drivers. We propose an EVR that rejects such usage of speculative benefits in determining acceptable safety, especially if such an approach is not disclosed to public stakeholders.

ADDRESSING PUMA CONSTRAINTS

Design engineers might not be in a position to change high-level management policies such as funding lobbyist pressure for prohibiting municipalities from having a say in autonomous vehicle testing activities that take place on their streets. However, given that an organization is inclined to consider EVRs in its design process, design engineers can still contribute significantly to ethical outcomes in other ways.

A key character of VBSRs compared to more typical functional requirements is that they commonly have more to do with what is prohibited than what is allowed. From an engineering optimization point of view, they tend to be more constraints on optimization rather than the target of the optimization.

For example, some company might say their strategy is to “balance safety and performance.” A more ethical EVR would be to “maximize performance subject to achieving acceptable safety,” with a VBSR of harm beyond a certain very low threshold being disallowed. Another EVR might be to avoid risk redistribution, with a VBSR prohibiting any increase in pedestrian harm compared to a relevant human driver baseline. So long as these VBSR safety thresholds are met, then performance can be optimized.

From a system engineering point of view, such an approach requires identifying and addressing what we shall call *Prohibited Utility Maximizing Actions (PUMAs)* at all levels of the organization. At the organizational level these traditionally take the form of regulations such as a ban on certain types of insider stock trading, illegality of faking federally mandated vehicle test results, and a requirement to maintain a specified type of insurance for operating vehicles on public roads. Other rules protect the integrity of markets by prohibiting price fixing and monopolies. While an organization might seek to increase profit by violating those PUMAs, doing so is prohibited by laws which top management of a firm is required to obey. The laws and regulations define the permitted scope of activity within which top management may pursue profit maximization.

The idea of a PUMA can be extended to engineering organizations as the basis of more actionable EVRs than general statements of ethical principles. A key type of engineering-relevant PUMA is specifically of the form of prohibiting an optimization strategy that might otherwise be attractive from a purely goal-seeking point of view.

Some examples of PUMAs from previously discussed ethical topics include prohibitions on:

- Over-use of designated at-risk areas and populations during public road-testing plan creation to reduce cost and testing time.
- Disabling the entire vehicle’s ability to operate if the manufacturer is losing money on software support only a few years into vehicle life.
- Using automation technology for which it is not possible to obtain specific safety metrics for comparatively rare objects and scenarios in support of making deployment decisions (e.g., wheelchair riders and pedestrians with darker skin tones).

Any requirements management process should be able to designate specific requirements as PUMAs rather than functional requirements. This should ensure that any violation or deviation requires an escalated review process beyond engineering discretion that might be applied for performance tradeoffs. PUMAs are not just nice to have – they exist to ensure that the design process is performed in an ethical manner, and as such they must have high priority in making design tradeoff decisions.

By way of comparison, EVRs are strategies that drive safety requirements. In contrast, PUMAs are business rules that prohibit some process and implementation strategies. PUMAs are complementary to EVRs.

ETHICS AND REGULATIONS

A design engineer might not really have ethical design first in mind when trying to debug some code or figure out how to get a system to pass a functionality test. Deadline pressure and an educational experience that is all about “getting it to work” do not often create an environment conducive to questioning the ethical implications of the design work.

A process aligned with IEEE 7000 can serve to incorporate the needs of various stakeholders, identifying EVRs applicable to the system. A system engineering process can create PUMAs and VBSRs as part of an ethically aligned process. Those make ethical issues actionable for design and test engineers.

The bigger question is whether corporations will actually follow IEEE 7000 processes, which are not mandated by law.

Despite corporate messaging that “safety is #1” and similar slogans, corporations are highly incentivized to optimize for goals that might be at odds with a robust set of ethical norms. This is not to

say that any particular company is purposefully unethical. Rather, if highly motivating incentives to erode ethical behavior exist, it is unrealistic to assume that they will have no effect on behavior.

Corporate officers have a fiduciary duty to maximize the value of their investor's stock. That mandate only considers ethical behavior as a secondary consideration – if even that. Indeed, corporate officers might be seen to have a duty to ignore the concerns of other stakeholders if doing so will maximize shareholder value.

In the final analysis, PUMAs traceable to regulatory pressure or government mandates are the primary means of incentivizing ethical corporate behavior. (There might be some social pressure, but success of such campaigns is the exception rather than the rule, and corporations are unlikely to react in any substantive way until such a campaign is launched and gets widespread enough traction to affect the bottom line.)

As a secondary layer of defense, in the automotive industry one might think that pressure from regulators and the insurance industry will force ethical outcomes. Sadly, this is far from the case.

The insurance industry can make a profit so long as they can predict eventual losses with some accuracy. A higher amount of harm caused by a particular technology equates to higher insurance premiums. So long as they can charge enough to issue policies, their business model is intact. While there are indirect incentives to advocate for safety of automated vehicle features, the insurance industry is not in a position to ensure ethical deployment of the technology.

This means that the only pressure for top managers in a company to act ethically comes from law and regulations. In the best case the top managers strongly want to act ethically, but they need air cover from laws and regulations to avoid being blamed for reduced fiscal performance for having done so.

While the risk of liability lawsuits is often stated as a deterrent force for bad actors, law is famously disjoint from ethics. Moreover, the extremely high stakes of the autonomous vehicle industry of billions of dollars chasing a trillion-dollar market can easily reduce the occasional payout of a few million dollars as victim compensation to being a mere cost of doing business.

While regulation cannot directly enforce ethical behavior, it can incentivize it by setting rules of engagement for competition between companies that reward ethical behavior and discourage unethical

approaches to maximizing profitability. This is a somewhat different dimension than the more usual discussion concerning whether governments should mandate particular safety standard conformance.

A critical service that regulators can provide is ensuring that different stakeholders have a voice in key issues such as risk exposure due to public road testing, lifecycle support requirements, and deployment governance. There are three alternative regulatory structures that might be taken.

The first is a self-regulatory approach, which is the one currently in place. That is not going well from an ethical point of view.² For example, a newly passed law in Pennsylvania prohibits cities such as Pittsburgh and Philadelphia from having any substantive say in where and how autonomous vehicle test platforms are operated.¹⁰ Municipalities are not even permitted to prohibit testing of immature automation technology in active school zones. The situation elsewhere in the US is not much better. In large part this is a direct result of successful autonomous vehicle industry lobbying for state laws that specifically prohibit local authorities from having a say, known as a municipal preemption clause. We can expect that the continuation of the self-regulation approach will continue to produce what amounts to a wild west of autonomous vehicle testing and deployment – until some adverse event too big to ignore forces action from regulators.

A second approach is using a system of recalls to apply regulatory pressure after harm has been done. This approach is starting to be used for vehicle automation technology at the US federal level, but with extremely limited effect. One problem is that significant harm might be done to many people before it becomes obvious that mandated recall is required. Another problem is that such an approach is unlikely to address many ethical concerns such as disproportionate harm caused to vulnerable communities. At best, a program of ad hoc reparations might be passed to address ethical lapses that detract from social justice (but society has not yet worked through how such a program might be implemented—or whether one should be implemented at all).

A third approach that avoids highly invasive regulatory oversight of the design process is a version of self-regulation that requires the industry to follow its own industry standards for safety and ethical behavior, including UL 4600 and IEEE 7000. While those standards do not force complete public reporting, coupling those obligations with an

independent auditing function and penalties for non-compliance could supply a combination of pressure and protection to encourage top managers to expend resources to fulfill safety and ethical requirements.

It hardly seems unreasonable to require an industry to follow its own self-written technical standards. Nonetheless the vehicle automation industry routinely pushes back against any such requirement.

An additional function that might be performed by regulators is setting a level playing field for establishing safety goals and defining a core set of PUMAs that apply equally across the industry. As an example, while the sound bite of “safer than a human driver” sounds appealing, defining and measuring that invokes incredible amounts of detail.²

A core set of PUMAs might also help put in place guardrails to encourage ethical behavior by companies engaged in the high-stakes race to deploy highly automated vehicle features. Example PUMAs that might help include:

- Testing cannot take place without reasonably actionable advance notification to local residents of times and places testing will occur.
- Vehicle safety cannot be degraded below that of a comparable model vehicle that does not have automated features if ongoing manufacturer support is terminated for any reason.
- Deployment cannot be performed without notifying stakeholders of a key set of decision-making process metrics according to a publicly disclosed decision-making process.

The point of having such regulations is to generate a level playing field so firms implementing ethical concerns as part of the design process are not at a competitive disadvantage. Such regulations should also help make the public reports of different firms within an industry more directly comparable. Comparability will assist greatly with social pressure exerted to consider ethical factors more generally.

Additionally, if including specific classes of stakeholders is mandated, there is less risk of overlooking an important class of stakeholder.

The least intrusive government response might be to require that a firm complete the IEEE 7000 process and publish the results for public viewing. Interested members of the public can be expected to perform the first-tier auditing function. Concern over adverse publicity will motivate a responsible firm approach. (IEEE 7000, at p. 27, recognizes that “[t]he success of a system can depend on indirect stakeholder opinions,

which can shape public opinion.”) With a regulation in place, it becomes in the interest of stockholders that the IEEE 7000 process be completed, and its results taken seriously, by top management.

CONCLUSION

Some law or regulation is needed to operationalize the use of IEEE 7000 to identify steps needed for the ethical development of a given new technology. Fiduciary decision-making at firms, standing alone, does not contain proper incentives for ethical behavior. Evidence from the vehicle automation industry’s track record continues to accumulate that supports the need for government involvement.

Government specification of certain standard safety-relevant performance metrics, identification of stakeholders, and certain social justice EVRs—perhaps in the form of mandated PUMAs—will facilitate comparisons among AV industry players and will create a more efficient internal IEEE 7000 process for companies and their design teams.

Regulatory processes are notoriously slow. In the meantime, vehicle automation design teams can get started establishing a set of EVRs, VBSRs, and placeholder PUMAs that apply to both design and validation activities to continue their journey creating ethically aligned, safe vehicle automation technology.

REFERENCES

1. IEEE 7000-2021 Standard addressing ethical concerns during system design.
2. P. Koopman., *How Safe Is Safe Enough? Measuring and Predicting Autonomous Vehicle Safety*, September 2022. ISBN: 979-8848273397.
3. Widen, William H., *Highly Automated Vehicles & Discrimination Against Low-Income Persons* (January 24, 2022). University of Miami Legal Studies Research Paper No. 4016783, North Carolina Journal of Law and Technology, Vol. 24, No. 1, 2022, <http://dx.doi.org/10.2139/ssrn.4016783>.
4. NHTSA, *Vehicle Data Privacy*, <https://www.nhtsa.gov/technology-innovation/vehicle-data-privacy>, accessed Apr. 27, 2023.
5. SAE Standard J3018_202012 *Safety-Relevant Guidance for On-Road Testing of Prototype Automated Driving System (ADS)-Operated Vehicles*.
6. U.S. Dept. of Trans., *Areas of Persistent Poverty & Historically Disadvantaged Communities*, <https://www.transportation.gov/RAISEgrants/raise-app-hdc> (last visited Jan. 23, 2023).
7. Colias, M., [“Americans Are Keeping Their Cars Longer, as Vehicle Age Hits 12 Years,”](https://www.wsj.com/articles/americans-are-keeping-their-cars-longer-as-vehicle-age-hits-12-years-11613482000) Wall Street Journal, June 14, 2021.
8. Kopetz, H. *An Architecture for Safe Driving Automation*. Ch. 4, *Principles of System Design*. Springer Lecture Notes on Computer Science (LNCS) Vol. 13660 (forthcoming Spring 2023).

9. P. Koopman, *The UL 4600 Guidebook*, ISBN: 979-8365303065, 2022.
10. Pennsylvania Act No. 2022-130, <https://www.legis.state.pa.us/cfdocs/legis/li/uconsCheck.cfm?yr=2022&sessInd=0&act=130> (last visited Jan. 23, 2023)

Philip Koopman splits his time between teaching safety critical embedded systems at Carnegie Mellon University and helping companies around the world improve the quality of their embedded system software. He was the lead technical author of the UL 4600 standard, and authored the book *How Safe Is Safe Enough? Measuring and Predicting Autonomous Vehicle Safety*. Contact him at koopman@cmu.edu.

William Widen is a Professor of Law at the University of Miami School of Law and an elected member of the American Law Institute. A graduate of the Harvard Law School, he is a corporate lawyer by training. His current research focuses on regulation of autonomous vehicles. Contact him at wwiden@law.miami.edu