Autonomous Vehicle Fleets as Public Infrastructure

Thomas Krendl Gilbert, Postdoctoral Fellow, Digital Life Initiative, Cornell Tech
Roel I.J. Dobbe, Assistant Professor, Faculty of Technology, Policy and Management, Delft University of Technology

Abstract

Autonomous vehicles (AVs) are increasingly recognized as a public problem. As they comprise a technological disruption of unresolved scale and complexity, it will be up to the public to decide what hoary terms like trustworthiness, legitimacy, and safety amount to in the context of automated transportation. But where exactly is this public? How can it be expected to emerge as public space itself is contorted by the modeling decisions and technical interventions at stake? And is it possible to mobilize this public more proactively? Drawing on 50 semi-structured interviews with technical experts across different stages of AV development, we focus attention on the specific *normative indeterminacies* introduced by this emerging technology. These indeterminacies correlate with particular public problems, which we argue are likely to manifest sociotechnically in the form of *moral crumple zoning*, *rubblization*, and *jayification.* In conclusion, we return to the philosophical theories of John Dewey and James Grunig to articulate strategies by which experts could engage the publics that may emerge in response to risks introduced by AVs, whether by listening to their feedback, aiding their organization, or representing their interests.

Introduction

The promise of 'autonomous vehicles' (AV) to redefine public mobility renders their development political across a variety of stakeholders. This politics may not be obvious.

On the one hand, with their ability to optimize the local safety and efficiency of individual vehicles, AVs promise to make individual transportation more predictable and reliable. Trips that people find too tedious to make could be made into trips worth taking, and as this change is reflected through the broader population, it has the potential to fundamentally change the relationship consumers have with transportation. AVs also make it possible to centralize and coordinate the routing of vehicles, thereby alleviating traffic congestion. Such works represent only the beginnings of what could be possible. Centralized route planning could allow load-balancing between routes on the scale of cities, the predictive placement of vehicles for the purposes of ride-sharing, special routing considerations for emergency vehicles, and the management of interactions between these considerations.

We claim these possible societal benefits, as well as the possible trend towards more centralized and concentrated forms of control and power to reshape public space, render the emergence of AVs a *public* problem. This means that it requires careful deliberation and consensus-making across societal stakeholders about what goals and constraints AVs should adhere to, now and in the future, and how these should be safeguarded across the technology, the traffic infrastructure, as well as its governance and institutional environment.

In this paper, we start sketching the public nature of the introduction of AVs and the need for such integral coordination and specification. To arrive at this sketch, we build on the recently developed *Hard Choices in Artificial Intelligence* (HCAI) framework, mapping emergent dilemmas and choices arising in AI system development to current and emerging forms of sociotechnical politics in particular application domains.[1]

The central dilemmas are explored across stakeholders engaged in AV development and governance. We examine the emerging regulatory landscape of AV development, based on 50 semi-structured interviews with researchers in AI theory, human factors, and AV policymakers.

[1] Dobbe, R., Krendl Gilbert, T., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*, 103555. https://doi.org/10.1016/j.artint.2021.103555

To our knowledge, this comprises one of the first qualitative datasets of insights and expert judgment from every stage of AV development, from design to training to real-world physical deployment.[2]

Based on this analysis, we identify four associated challenges for the responsible public development and integration of AVs. First, AVs are disrupting legacy processes for vehicle safety certification. We are witnessing regulatory capture as AV companies hire federal and state contractors to ensure their design certifications meet legacy thresholds for liability. Companies now craft their own *Operational Design Domains* to meet proprietary definitions of road features (streets, lanes, city regions) that purport to be technically safe, without a more complete sociotechnical account of the driving and road environment. Second, traditional forms of oversight are now being placed downstream of AI-enabled techniques such as *Adversarial Learning* to make AVs robust to particular traffic dynamics and forms of signage that are easier to model. This leaves pedestrians vulnerable to the system specification and potentially compromises public criteria for safety and fairness. Third, there is the metaphorical frame through which AVs are currently narrated and marketed to potential consumers, as private companies, consulting firms, and municipal entities craft outreach surveys as they see fit to shape the types of demand that suit their own organizational priorities. This in turn risks a violation or incomplete safeguarding of public concerns.

By recruiting the HCAI framework to map concrete AV development choices to current and emerging forms of sociotechnical politics, we start painting a more granular picture of what central questions arise for understanding AVs as a public problem. In dialogue with the existing literature, we summarize three dimensions of AV politics: *jayification* (which places certain mobility stakeholders "out of scope"), *rubblization* (which remakes the road environment to suit convenient technical thresholds for model robustness), and *moral crumple zoning* (which allocates responsibility for accidents to the most proximate human).

Integral across the other challenges and of a more fundamental nature, is the need for *sociotechnical specification,* which we defined in an earlier paper as "the process of facilitating the different interests relevant in understanding a situation that may benefit from a technological intervention."[3] As we will show, the common thread across all three issues is the need for expert professions to take a more active role in organizing and representing distinctive "publics" as they come into being through the development of automated transportation capabilities.

---

[2] Jack Stilgoe, as part of the "Driverless Futures?" project, has recently published work resulting from "50 interviews with self-driving car developers, researchers and policymakers". While the content and themes of these interviews focused on AV safety criteria, rather than technological and institutional development of AVs as a whole (as we do), we look forward to engaging this work more closely in a future draft of this paper, in particular on the question of incorporating perspectives from the public at large. See Stilgoe, Jack. "How can we know a self-driving car is safe?." *Ethics and Information Technology* (2021): 1-13.

[3] Dobbe, R., Krendl Gilbert, T., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*, 103555. https://doi.org/10.1016/j.artint.2021.103555

In our treatment of the public we follow in the footsteps of, who stated that "[t]he public consists of all those who are affected by the indirect consequences of transactions to such an extent that it is deemed necessary to have those consequences systematically cared for."[4] The current lack of awareness and understanding of AVs as a public problem, and the associated institutional void, leave the underlying normativity of AVs "up for grabs" and subject to manipulation in a way that comes at the direct expense of human passengers, potential commuters, and public infrastructure.

Despite the common label of AVs as "autonomous", they will be shaped by human interests and expectations. Their status as an emerging form of public infrastructure must be decided through ongoing normative deliberation and an institutional landscape which sustainably represents and safeguards the public interest. Ignoring the nature of AVs as a public problem may not only generate detrimental risks to public safety; it could also render unsuccessful the project of AVs and stand in the way of more promising futures for mobility.

The Problem of Public Space

Normative critiques of artificial intelligence (AI) systems have become commonplace. For example, we have now seen at least two waves of "critical algorithm studies". The first presented critiques of the big data paradigm, focused on exposing its spurious claims to eliminate human biases from historically collected data.[5] A second wave has zeroed in on particular technical failures related to traditional categories of discrimination based on race[6], class[7], and gender[8]. More recent work is drawing attention to the distinctive political economy of formal algorithmic systems, including critical investigations of mechanism design[9] and renewed attention to the legal tool of antitrust.[10] Even so, emerging normative questions related to the ethical, legal and

---

[4] Dewey, John. *The Public and its Problems*. Pp. 15-16.

[5] See Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *California Law Review* 104, no. 3 (2016): 671-732. http://dx.doi.org/10.2139/ssrn.2477899; O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books, 2016; Pasquale, Frank. *The Black Box Society*. Cambridge, MA: Harvard University Press, 2015.

[6] Benjamin, Ruha. "Race after Technology: Abolitionist Tools for the New Jim Code." *Social Forces* 98, no. 4 (2020): 1-3. https://doi.org/10.1093/sf/soz162; Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

[7] Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press, 2018; Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books, 2019.

[8] Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." arXiv preprint arXiv:1607.06520 (2016).

[9] Viljoen, Salomé, Jake Goldenfein, and Lee McGuigan. "Design choices: Mechanism design and platform capitalism." *Big Data & Society* 8.2 (2021): 20539517211034312.

[10] Newman, John M. "Antitrust in digital markets." *Vand. L. Rev.* 72 (2019): 1497.

societal implications of AI systems tend to be addressed in separate methodological or disciplinary silos whose categories are fixed.

The prospect of partially- or fully-automated vehicles (AVs) has been addressed in a similar fashion. Originally, much ink was spilled on the "ethics" of self-driving cars cast in terms of trolley problems--a contrived situation in which an impartial observer must decide whether the car should run over a group of people by default, or a single person intentionally. This is often framed as a choice between rival normative commitments to deontological or utilitarian codes of ethics. Trolley problems, however, have been critiqued for their narrowness, contrived nature, and inappropriateness, particularly with respect to the more concrete political questions posed by automating critical components of the transportation system. Researchers have highlighted concerns about the safety of AVs by drawing attention to user interfaces, multi-modal interactions, and simulation parameters rather than the exclusively formal modeling assumptions at stake in trolley problems. Additional work on the potentially-discriminatory biases of computer vision and monopolization of the road environment[11] by automated fleets has echoed critiques of AI systems in general.

At present, however, there is little work understanding how AI systems or AVs reshape public space and functions. This is unfortunate, as the historical record indicates that this is a more basic problem than the specific transformations of social categories engendered by new technology. Peter Norton's illuminating history of the automobile's introduction to American streets reveals how new social formations were able to be mobilized for the first time in response to the accidents and congestion generated by cars.[12] The resultant stakeholder categories (pedestrians, safety reformers, police, street railways, downtown business associations, traffic engineers, and motordom) were largely contingent on the urban, demographic, and economic concentrations of particular cities. Moreover, the boundaries of these categories were not consistent or determinate, and were in fact structured patterns of activity that active road users came to interpret as semi-exclusive identities in the context of the particular vulnerabilities introduced by cars. In large part, the historical narrative of streets' appropriation by vehicles was about the fragmentation of the public interest, as identities first congealed, and were sorted, according to a new understanding of public space that was then materially enacted in the demarcation of roads into car lanes, (narrowed) sidewalks, standardized signage, and vehicle-exclusive highways.

Similarly, we are today grappling with the broader politics of AV development, and their safety and role in reshaping mobility and public space.[13] In particular, a more integral perspective

---

[11] Andrus, McKane, et al. "AI development for the public interest: From abstraction traps to sociotechnical risks." *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2020.
[12] Norton, Peter D. *Fighting traffic: the dawn of the motor age in the American city*. Mit Press, 2011.
[13] Jack Stilgoe's recent work is an excellent example of normative engagement with this problem: "Rather than taking the technology as fixed and looking to plug the deficits of law or public understanding that are imagined around it, policymakers should instead see self-driving cars as an opportunity for more active

linking the problem of public space to the design challenges and technical affordances of AVs is missing.[14] Outside a somewhat small community of civil society advocates[15] and piecemeal academic discussions[16], there is to our knowledge no concerted effort to make sense of and strategically organize the sociotechnical horizon of transportation that automated mobility makes tractable.

Such an agenda has at least been sketched for AI systems, both in an abstract sense and more recently with application to AVs. For example, we can learn from a long tradition of system safety approaches, which have been developed for software-based automation in safety-critical fields. While recent connections have been drawn between system safety and the particular affordances of AI systems[17], these lessons still have to be translated to specific domains newer to automation, including AVs and road traffic. In addition, Goldenfein et al. have recently proposed the "handoff model" in an effort to represent alternative visions of how AVs could be integrated onto roads, mapping onto different classes of actors who believe in and enact those visions over time.[18] This model emphasizes the sociotechnical reconfigurations of the driving environment, capturing how discrete driving tasks are "handed off" across both human and technical actors. These handoffs in turn generate qualitatively distinct normativities in the form of ethical mandates and propositions that are particular to human formations assumed to operate within each vision.

Inspired by this work, we believe the next step is a richer empirical understanding of the linkages between specific technical interventions and the reconstitution of public space in the context of AVs. Beyond tracing out the cohesive normative visions implicit in the work of current actors, as well as attempts to encode values directly into AV systems á la trolley problems, this would engage directly with the problem made stark by Norton's research: where and how the public interest manifests in distinct stages of technological disruption as it proceeds.[19]

---

engagement in the shaping of technological systems, prioritizing social learning and knitting self-driving cars back into their social worlds". See Stilgoe, Jack. "Machine learning, social learning and the governance of self-driving cars." *Social studies of science* 48.1 (2018): 25-56.

[14] Even more recently, Stilgoe and Badstuber have emphasized that engagement with the public is necessary because AV technologies are, as a fact of the matter, developed in public. See Stilgoe, Jack, and Nicole Badstuber. "Democratising Driverless Futures: Five Lessons for Public Dialogue on AVs." *Automated Vehicles Symposium*. Springer, Cham, 2019.

[15] Rode, Philipp, et al. "Towards new urban mobility: the case of London and Berlin." (2015).

[16] Alessandrini, Adriano, et al. "Automated vehicles and the rethinking of mobility and cities." *Transportation Research Procedia* 5 (2015): 145-160.

[17] Dobbe, Roel. "System Safety and Artificial Intelligence". To appear in Oxford Handbook on AI Governance (2022)

[18] Goldenfein, Jake, et al. "Through the Handoff Lens: Competing Visions of Autonomous Futures." *Berkeley Tech. LJ* 35 (2020): 835.

[19] The problem of evaluating alternative strategies for engaging the public itself in the context of AV design choices has been recently explored in Stilgoe, Jack, and Tom Cohen. "Rejecting acceptance: learning from public dialogue on self-driving vehicles." *Science and Public Policy*, 2021, 00, pp. 1-11.

Normative indeterminacy and the Hard Choices framework

To make sense of the situations in which transportation criteria now appear as public problems, we mobilize the Hard Choices in Artificial Intelligence (HCAI) framework. HCAI articulates the most fundamental AI design problems through the lens of *normative indeterminacy*. AI applications introduce normative indeterminacy when they operate or intervene at scales that were previously inaccessible to humans. The effects of these interventions are defined not just by uncertainty (which in principle could be modeled), but by a fundamental lack of standards or guidelines for how to evaluate the performance of the system in question. Many of the capabilities of future, highly-advanced AI systems will be defined more by normative indeterminacy than uncertainty.

While a full presentation of HCAI is beyond the scope of this paper, we highlight two key claims it makes to address the existential problem of normative indeterminacy. First, AI development must be reconceived in terms of the multiple points of encounter between system capabilities and normative indeterminacies. Collectively, these points are referred to as the *sociotechnical specification* of a given AI system. An exhaustive appraisal of sociotechnical specification entails a new vocabulary to make sense of salient indeterminacies in the context of technical design decisions, constituting a reciprocal relationship between system development and governance. In the next two sections, these elements are presented in terms of the *featurization*, *optimization*, and *integration* of the system in question. Second, developers must take on new roles that are sensitive to feedback about how to manage these indeterminacies. This requires communicative channels so that stakeholders are empowered to help shape the criteria for design decisions. HCAI specifically refers to these channels through the need to accommodate *dissent* as a particular mode of out-of-distribution feedback from stakeholders, motivating the need for alternative sociotechnical specifications.

In summary, HCAI catalogues the indeterminate design situations in which public problems are able to emerge, and thereby suggests strategies for how to nurture rather than stunt the public's maturation. Both these claims are of direct and immediate relevance to AVs.

Empirical methodology and interview procedure

This paper is based on over fifty (50) interviews conducted with researchers, manufacturers, and regulators in the automated vehicle industry. Where possible, interviewees were selected based on professional maturity (~5+ years working in their current position), although younger interviewees were chosen if they were outstanding members of their respective field. Interviews were conducted in person whenever possible, either at the subject's place of work or at a leading

conference venue such as the Transportation Research Board annual meeting; about half were conducted virtually over Skype or Zoom, in part due to the COVID-19 pandemic.

Interviews averaged sixty (60) minutes in length, but could go up to two hours depending on the range of topics discussed. After examining all interviews, reflecting on key themes, and coding them according to the scheme described below, eleven (11) of these were selected as the basis for discussion and analysis. These interviews were chosen for their subjects' particular degree of expertise within relevant topic areas (respectively human factors research, AV survey experience, and machine learning engineering), the depth of awareness and ability to articulate topics of public concern, and the variety of experiences across state, federal, non-profit, and profit-based organizations that work in the space of AV development and regulation.

Most interviews followed a similar script, split into three parts. Participants were first asked about their career trajectory, including their motivations for working on automated vehicle development, key steps in their professional journey (undergraduate and graduate training, previous jobs, summation of their current position), and how they expect their own role to evolve as AV development and deployment accelerates. Second, participants answered a series of questions about their views on salient design considerations of AVs. This did not include trolley problems (to the surprise of some participants!), and instead focused on design challenges at different scales of normative abstraction. Typical topics included the key technical components of a successful driving experience (e.g. passengers, emergency takeover drivers, pedestrians, other cars), changes in infrastructure necessary to handle the AV transition successfully (including smart roads, scalable ride sharing, 3D roads, protected neighborhoods, policy incentives), and their own views on current or prospective public-private partnerships. These topics and particular questions were adapted to match the expertise and perspective of particular interviewees. Third, participants were asked about how they perceived the reconstitution of values (e.g. safety, security, privacy, convenience, comfort, transparency, explanation) through the introduction and widespread adoption of AVs. These questions drew heavily from the tradition of speculative design, and were chosen to interface thematically and substantively with ideas in the long-term problem space of AI governance and safety.

The purpose of these questions was to capture participants' engagement with public problems. These could include their own mental models of "the public" as a discrete political or social agent, the nature and size of public space, and the prospective status of AVs as a public good. By tracing the dispositions of participants towards various forms of indeterminacy in the domain made manifest by the prospect of AVs, we could indirectly construe the image of the public at stake in a given technical or regulatory approach. Often this was achieved by asking respondents to weigh in on a topic that, on its surface, did not relate to the public at all. For example, participants were asked to present a "principled compromise" between preserving the privacy of individual AV passengers--including protection and ownership rights of their own data--and

optimizing the safety of AVs so that as many crash scenarios are avoided as possible. The purpose of this question was to encourage the participant to articulate what, in their own eyes, was the public interest with respect to the incommensurate values of privacy and safety. In that case, the participant was asked to air their own views as a member of the public at large, rather than with reference to their own preferences or personal sense of utility. Participants were sometimes asked to distinguish their own views from what they imagined most other members of the public to want; this allowed us to distinguish between how the participant discerned the public as a representable social object vs. the participant themselves as a prospective member of the public in question.

Interviews were coded on the sentence level according to preference for a certain manner of reflecting on, representing, or otherwise engaging public problems. Coding categories were derived from the typology of the HCAI framework. More specifically, when respondents described a source of indeterminacy in a given computational model, user interface, rhetorical frame, or some other technical aspect, these were coded according to the following types:

- Featurization: absence of criteria for evaluating the environmental components of a given formal model
- Optimization: absence of criteria for evaluating what “good” performance means of a particular AV system
- Integration: absence of criteria for evaluating the interface between the AV system and external conditions

By tracing how our respondents referred to different sources of indeterminacy in response to prepared questions, we were able to correlate attention to particular public problems with distinct modes of technical expertise. This in turn indicated to us the emerging role of various elite professions in defining, probing, and communicating with the situational publics likely to emerge as AVs are further developed and deployed.

We note that this coding is provisional, to be refined both through further examination of additional interview content as well as follow-up data collection. We are also curious to conduct follow-up interviews with particular respondents to see if or how their mental models of public problems adjust over time as AV development continues to rapidly change. For now, we present our research findings as a snapshot of the present sociotechnical landscape of AV regulation, both in an empirical sense and as a guide for future normative interventions.

Investigating Indeterminacies

This section outlines major qualitative insights from interpreting and coding interviews conducted with human factors researchers, AI theorists, and regulators on the problem space of

autonomous vehicles (AVs). Below we list prominent forms of regulatory indeterminacy in how AVs are currently designed and marketed, and characterize the politics for each based on the emerging conflict over how to resolve this indeterminacy through alternative (and competing) definitions of the problem. We also map these forms of uncertainty onto the sources of indeterminacy in the HCAI framework. A tabular depiction of our findings is presented in Table 1, whose terms are progressively referred to throughout this section.

Many interviewees struggled to articulate a vision of the public as dynamic and active, rather than static and passive. The "public" was something to be represented, measured, marketed to, or lobbied, rather than something to be attentive to, nurtured, or obeyed. Distressingly, this attitude was shared as much by federal and state regulators as by designers, though the former are nominally charged with representing the public's interests.

HCAI component	Source of Indeterminacy	Technical example	Public problem	Sociotechnical example
Featurization	Operational Design Domain	Level 3+ autonomy	Liability	Moral crumple zoning[20]
Optimization	Road signage	Adversarial learning	Damage to infrastructure	Rubblization[21]
Integration	Marketing frame	Consumer surveys	Mobility	Jayification

Table 1: The politics of automated vehicles, beyond safety.

Indeterminacy 1 - Decoupling autonomy and responsibility

The classification scheme presented by the National Highway Traffic Safety Administration ranks autonomous vehicles on a scale from Level 1 (full human driver control) to 5 (full vehicle autonomy). This is meant to serve as a regulatory signpost for the current and eventual technical progress on AVs, i.e. that for each specified degree of technical autonomy there is a corresponding regulatory standard that indicates safe deployment conditions in a well-defined way. Level 1 autonomy broadly refers to automated systems that have become more common in cars since the early 2000s, such as adaptive cruise control that adjusts the speed of the car to ensure a safe following distance based on the perception of external road features. Level 2 technologies include lane centering that also automate steering functionality, and have been increasingly implemented throughout the 2010s.[22] Tesla Autopilot, first technically enabled in 2015, remains classified as a Level 2 autonomy system; Waymo prominently tests and markets AVs with Level 4 autonomy, which has offered commercial services since December 2018 in Phoenix and since August 2021 in San Francisco.

[20] Elish, Madeleine Clare. "Moral crumple zones: Cautionary tales in human-robot interaction (pre-print)." *Engaging Science, Technology, and Society (pre-print)* (2019).
[21] https://en.wikipedia.org/wiki/Rubblization
[22] As discussed in my interview with Chuck Green.

Considered from a sociotechnical standpoint, there is an indeterminate relationship between the technical specifications underpinning each of these levels of autonomy and the equivalent degree of attention and responsibility expected of human drivers. This tension is crystallized at the *Level 3 stage*, i.e. a mode of driving in which nominally the vehicle is able to handle all driving capabilities under specified conditions, but in case of an emergency the human driver must take over and retain responsibility for any resulting accidents and damage. In this case, an "emergency" is typically defined as a scenario the system cannot navigate[23], requiring a notification to the human driver whose limit of reaction time for taking action is traditionally specified by human factors.

The viability of Level 3 systems is most prominently discussed and debated in terms of guidance from the human factors community. But in fact, this research reveals the normative incommensurability between interpretations of the street as built for different forms of agency: drivers (who are thus responsible for resulting damages, as they are interpreted as "misusing" driver assistance tech) or AV designers and manufacturers (who would thus bear the cost of Level 3 accidents and other unexpected externalities of the incomplete specification, as they are interpreted as "negligent" with respect to deployment conditions).

> So essentially a Level 2 system a manufacturer is making the same sort of statements in the owner's manual that you're used to them making today. Which is, if you operate this vehicle according to the way that you were trained how to drive, then the vehicle will do what it is supposed to do. Whereas, receptivity is something totally different. Receptivity, Level 3 is basically you are telling a person, you are not driving. You are watching TV and email. We will let you know if we need to drive. It's our responsibility... that includes not just an alert because something happens, but the system has to also detect when it's no longer able to do its job. So, if the weather changes too much or if it is supposed to only work at a highway and the freeway is ending. The system has the role of telling the user that it's no longer going to be able to perform like it should and the user needs to take over and one of the most interesting topics is, first off, a lot of people are not sure that user can even play that sort of role um will they ever be able to take over if they weren't situationally aware to begin with, but the second thing is there's a lot of discussion about um simply it's often just talking about how long...Clearly there's a human factors limitation, you can't ask a person to become situationally aware and do all the jobs of driving instantly. But clearly there's some time, right. I remember I was in a meeting once and there's a really perceptive, another expert there and he compared it to catching and releasing a fish. He said, if you intend to catch and release a fish. He said that if you just take the fish off the hook and throw it into the water, he said the fish will die, that's not

[23] https://www.caranddriver.com/features/a15079828/autonomous-self-driving-car-levels-car-levels/

> how you do catch and release. The whole idea of catch and release is not to kill fish. But if you just throw it into the water, if you take the hook out of its skull, it's going to die. And the reason is because of the shock of the change of its environmental conditions. He said the way that you do catch and release is you take the fish and you hold it and you put it into the water with your hands on it and then you wait until it begins swimming and then you let go. And he said that's the way that the fish will stay alive. So, he said in this way with the automation, you can't just throw a person and say, oh it's time for you to drive, you know, it's not about whether it's 2 seconds, 5 seconds, or 20 seconds. He said it's really about giving the opportunity to start swimming down that highway and then once a person basically has started swimming it started becoming situationally aware of the environment, getting comfortable with the steering and the other controls from the vehicle, and so he said it's really not certainly a matter of time, actually more than that. (quotes from interviews are edited for readability)

The taxonomy of autonomy levels created by NHTSA is a deductive attempt to fit specified human factors roles into indeterminate sociotechnical positions, and is not a substantive appraisal of how technical capabilities map onto reality. At best, for the human factors community, it is an "interesting research question whether humans are going to be capable or very good at performing the roles that they've been allocated" in a Level 3 system.

Indeterminacy first manifests in the need to define what "safety" means in the context of a partially-automated driving system:

> Interviewee: There was a NHTSA effort I was involved in called Automated Vehicle Research, it was the name of the contract. It was industry consortia contract, and there we used a system safety technique called STPA or System Theoretic Process Analysis, I think and as part of that technique the first thing you do is define what safety is and in this case, it's trying to reduce the likelihood of what they call a loss. So, a loss can be any injury or fatality for people and it also includes property damage.
>
> Interviewer: Okay, so it's all encompassing in that sense.
>
> Interviewee: It's actually, in many ways these definitions of safety are actually narrowly defined. A lot of human factors research really talks a lot about trust and comfort and this kind of research, it's not about trust or comfort. It's literally about the likelihood of loss. It's about reducing the likelihood of loss.

Safety has different meanings depending on the terms under which a safety concern is identified. Safety benefit, for example, refers specifically to the demonstrable reduction of crash risk due to evidence compiled from case studies, either directly on public roads or on test tracks under controlled conditions. Safety criticality, meanwhile, refers to a system feature whose failure would generate a hazard, i.e. a possibility of loss for the system as a whole. Part of what makes automated vehicle systems difficult to evaluate is that they are able to be safety critical and simultaneously provide a safety benefit. For example, even a relatively simple Level 1 system like adaptive cruise control could generate longer following distances and reduce crash risks on average, but also introduce safety criticality if it fails or is misused by the human driver.

This generates indeterminacy about the underlying constraints needed to ensure desired safety improvements. The combination of software vulnerabilities and human factors requires principled coordination of design, operation, governance, and institutional oversight, based on the designer's own judgment about the hazard's nature, severity, level of needed mitigation, and what it is possible to ask from a human operator in order to maintain system stability. This problem is wholly distinct from attempts to make the system safe or explainable via formal modelling alone, as the latter ignores the need to define acceptable model failures and worthwhile thresholds of risk:

> Systems of all kinds, especially automation systems, have displays and the purpose of these displays is so that the user understands the mode of the system, what mode it's in and so they can help predict what actions the system might take that can understand what's happening inside the system. It's to reinforce and help build the user's mental model, and this is very important for a complex system like a nuclear reactor or driving automation system in a car or an airplane. These issues of a person knowing how the system works so they have a proper mental model of how system operates, understanding the current actual mode of the system, analyzing every possible way that a person's understanding of the system could go wrong and all the things that could happen is a great thing to do as engineers and as researchers, and it's certainly great to try and address as many of those things as you can, but it's also really important to collect as much evidence through research as you can. [This is] because human factors is still primarily an empirical field, and you might find that many of the things that could go wrong don't, and this is certainly important because when you're trying to produce a product in an industry, you could quickly make this product completely unaffordable. So, for example, requiring a heads up display in a driving automation system so that you can display all these things that the system knows without taking the driver's eyes away from the environment. Heads-up displays are very expensive, large ones that could do all that are very expensive. It turns out you don't really need that because there's all of these sorts of redundant

> channels, and sometimes it's more important to actually simplify these displays as much as possible because essentially looking at your system display is a secondary task. The primary task of the driver in a Level 2 system and the Level 1 system is to look out the window and so, you can go overboard imagining all the different ways and doing analysis to find all the different ways that a person is misunderstanding the system and end up producing this like, very rich display environment which is counter-productive.

The risk is that AV developers exploit this indeterminacy of the driving environment by redefining it as they see fit without taking responsibility for its material construction and acceptable behavior patterns. Corresponding strategies could include marketing AVs as "worth the risk", or suggesting publicly-funded renovations to road infrastructure to lift it up to the system specification. There is no default organizing principle for the *automated features* of the vehicle, the *ergonomic features* pertaining to the driver, and the *external features* of the road environment. There is an emerging politics on this spectrum of features that pertains to defining what this organizing principle should be or could be in order for the entire system to function well. Concretely, this manifests in terms of the *Operational Design Domain* of the automated vehicle, i.e. the features specified by the operator to be technically safe at a given level of operational autonomy.

> Now specifically, I think maybe one thing that will help is talking about the operational design domain. So, you can have just like all of the features, the level of automation, the operational design domain really impacts your visceral understanding of what the feature is or how to use it. For example, you can have a system which does the steering and does the longitudinal control with driver supervision, only does it parallel parking. So, that is technically an L2 feature...The work that I did on LAADs was primarily looking at an operational design domain of freeways which divided limited access and looking at speeds through the full range of those freeway speeds. So, heavy traffic down to stop all the way up to something in the order of 80 or 85 MPH. By the way, this was precursor research that eventually General Motors produced the Super Cruise feature or product and that has the same operational design domain.

The Operational Design Design (hereafter ODD) comprises the sociotechnical horizon within which the autonomous functionalities of the vehicle are claimed to be able to perform well. It is not a scalar function of performance, but an operational matrix and legal interface in which the incommensurable features of software, human, and road are resolved and sorted according to responsibility rather than formal model representation. It captures the practical significance of features for a given task, as well as who must account for them if something goes wrong.

The corresponding aspect of HCAI is *featurization*, the definition of the situation and context within which the AI system in question is intended to operate. Distinguishing the features that matter from those that do not remains the purview of human AI designers; a self-driving car cannot in any meaningful sense decide what objects to recognize or actions to take, and is merely attending to features that matter according to what it has been told.

The relationship between ODD and featurization is illustrated through the indeterminate concept of *receptivity*:

> In order to be Level 3, by definition, the manufacturer must promise that if the driver does not supervise the action of the feature, the performance of the feature in the environment, any environment that safety will be maintained, that is the role of the system is to perform object and event detection response and not the user. So, the driver in this case, falls back to where the user in terms of Level 3 is not, the role is not to monitor the system or the environment. The only thing in Level 3 that the human user is required to do is something that's called receptivity, be receptive. [...] So, like for example while we're talking now, you are not monitoring your mobile, you are not staring at it looking to see if anything is happening with it. However, if somebody calls you or sends you a text, your mobile phone will produce an alert and you are receptive to those alerts. So, if someone calls you, you will notice that that has happened and you'll be able to take the call. Now there's of course many new issues like, 'I'm not gonna do that right now, I'm busy,' blah blah. But it's just an analogy. So, [the] analogy is receptivity means you will notice the alert and you'll be able to take the call, whereas, in a supervision which is the Level 2 role it would be like, you'd be staring at your phone looking for a call that comes in, when that call might not necessarily have any ring or anything else, just maybe the call would come in. So, that's the analogy that I would use. Level 3 basically means you can perform according to [the] manufacturer's directions, you can actually perform non-driving related activities. So, you could read, you could do email or text on your phone, you can even watch a video as long as you were attuned to whatever the alert mechanism was...The difference between Level 2 and Level 3 is actually all about the proper use of the system, or what I'd like to say is **the promise the manufacturer makes, it doesn't**. I mean there are implications on the technology but the technology itself is not part of the standard. (emphasis added)

Level 3 autonomy, in other words, is not at all analogous to Level 2 with more automated capabilities added. It is more analogous to Level 4, with the addition of a legal contract between the driver and manufacturer that the former will be alerted in case of emergency and is thereafter

responsible for whatever happens. Whereas Level 2 autonomy assumes that the automated capabilities are continuously supervised by the human driver, Level 3 autonomy assumes the driver is *periodically receptive* to alerts from the system to take over. This amounts to a distinction between a road environment defined by driver control vs. AV control, comprising incommensurable normative frames:

> Supervision means that the user is monitoring the driving environment, and the user is also monitoring the driving automation system, and the user is monitoring how the driving automation system is performing...Part of the supervision means taking over from the system [if] the user is not satisfied with how it's performing in the environment. So, supervision means you have to do all of those things all of the time. In the J3016 standard we talk about a concept called OEDR which is an acronym, Object and Event Detection and Response. So an object might be a vehicle or a pedestrian but it's also an event. An event could be the road is changing its character, like it's splitting or there's an exit. An event could also be nearby traffic is changing its behavior--it's slowing, it's speeding up, it's changing lanes. An event could also be that the weather is changing and now that it's raining, for example, it's no longer appropriate to use [the] automatic launch control system when the road is getting slippery. So, OEDR is more than just avoiding, say, an obstacle on the road. It's essentially assessing the complete character of the driving environment, assessing and predicting how the system is and might perform making judgements about whether to take over, whether to deactivate the system. So, that's continuous. Receptivity is only one thing, which is being able to determine if the system has issued its take-over alert, and then if so, being able to essentially begin driving after that. In theory you can have an alert that simply tells the person to begin supervising. So, maybe the alert would go off and then the person only needs to start watching. But oftentimes people think of request to take-over as being a request for the person to at least begin steering, if not also begin longitudinal control.

Receptivity entails a total handoff of the driving situation to the human operator. It is a microcosm of the fact that the transition from Level 2 to Level 4 is primarily rooted not in technical refinement of the system but in an altered social contract about who bears responsibility for its performance, based on the relative proximity of salient features to human awareness. Additionally, the capabilities of Level 4 cannot be understood in terms of scalar improvement from Level 3, but in terms of the AV manufacturer’s altered legal vector, in particular its willingness to be at fault for mishaps that affect system performance. There are completely incommensurable models of the human operator at stake in these distinct levels, from active monitoring of the entire driving environment, to ability and willingness to take over in case of emergency, to near-total passivity unless system override is desired.

The public problem of *liability* flows from the indeterminacy at stake in what features help comprise an AVs' ODD for a given level of autonomy. A more capacious ODD that includes more features such as lane-switching, longer-term horizons for motion planning, or the ability to coordinate entire fleets of driverless cars could only exist at higher levels of autonomy and may well amount to the AV firm taking responsibility for a much greater range of damages related to system performance. A more restricted ODD such as GM Super Cruise or Tesla Autopilot assumes Level 2 or lower autonomy, meaning that the human driver would be responsible for crashes in almost all instances. But in the case of Level 3 autonomy, an AV manufacturer could knowingly place features within the ODD that are difficult to model or control, yet find the driver responsible if they were in principle in a position to mitigate them in real time based on receptivity. If left unchecked, this strategy would leave human factors researchers in the unfortunate position of mitigating inattention and misuse under conditions of driver liability--in other words, aiding the design and implementation of a system that is able to reliably detect and report when it can no longer do its job. Instead, they could help hasten the emergence of automated systems for which AV firms themselves are liable.

In the domain of automated transportation, the key sociotechnical example of indeterminate featurization is what Madeleine Clare Elish calls *moral crumple zones*.[24] A moral crumple zone is the misattribution of responsibility for technological system failure to the nearest human operator, effectively shielding the system itself from liability. The indeterminacy of Level 3 autonomy leaves the underlying normative frame of AVs "up for grabs" and subject to manipulation in a way that may come at the direct expense of human passengers, potential commuters, and public infrastructure. Put differently, we interpret Level 3 autonomy itself as the "moral crumple zoning" of an automated vehicle's ODD, expanding its purview to include more features in the name of driver liability rather than technical feasibility.

[24] Elish, Madeleine Clare. "Moral crumple zones: Cautionary tales in human-robot interaction (pre-print)." *Engaging Science, Technology, and Society (pre-print)* (2019).

An example crash of an autonomous Uber vehicle.

More specifically, we can think of moral crumple zones as arising from development situations in which critical features of the road environment are put within the ODD in ways that are not clearly communicated to or understood by potential stakeholders, yet who are held liable for them. In an even more basic sense, these stakeholders would be unable to adjudicate between possible design choices. At some stage of development time, a compromise was struck between domain featurization and human factors, such that drivers ended up paying a political and moral price for making "mistakes" behind the wheel.

In a more abstract sense, the interface between ODDs and autonomy level determines AV accident liability as a public problem. This is an emergent form of sociotechnical politics that, to be avoided, must be replaced by more responsible and stakeholder-sensitive design commitments. With respect to human factors, there is a difference between tailoring the system to what is understood about the human on the one hand, and shaping human behavior dynamically alongside the system on the other (also known as operant conditioning), so long as this is done with the goal of system stability rather than general control of how a situation develops:

> On most human factors you try to understand human behavior and then you design a system that will sort of shape [itself] to that. In this case, we actually are shaping consumer behavior dynamically with this system. The driving attention system actually shapes customer behavior and it does it using principles from psychology. In particular, everybody's heard of B.F. Skinner and Behavioral Adaptation. That is the principle, it's officially called Operant Conditioning and that is essentially the principles we used. So, the system performs operant conditioning on the driver in order to mitigate misuse while driving.

> Behavioral modification is a very specific area. It's not a general thing. I know the word sounds general, but behavioral modification...operant conditioning as I understand it is, there is one specific behavior you're trying to change and so one of the specific things that are done to try and change that specific behavior. It's not a general thing, that I can generally change people's behavior in a general way to anything I want. So, for this stop sign example, I don't think the same kind of techniques, the operant conditioning techniques would work at all, operant conditioning techniques, they work overtime. It's like a practice, and the simplest kind of conditioning is like Pavlov's dogs. I mean, they didn't start salivating the first time he rang the bell, it's something that happens over time as training. That's why they're called conditioning, so it's not exactly parallel, you wouldn't use operant conditioning to affect the behavior of a person at a four-way stop. However, just characterizing the problem as, what can I do with my system to change the behavior of that person who's also in my system.

Indeterminacy 2 - Manipulating public anxieties

There have now been many prominent surveys of consumers' willingness to ride in or purchase automated vehicles. These range from personally-owned AVs,[25] AVs with the option of human driving[26], public roads that include driverless cars[27], driverless freight trucks vs. passenger vehicles[28], and many other topics.[29] The primary goal of these surveys--as expressed by the people who have designed and implemented them--is not to measure public opinion. Rather, it is to inquire into and actively shape it by encouraging sampled communities and demographics to reflect in a structured way on the conditions under which they would or would not see AVs as trustworthy.

> We are at an interesting place right now in the development, in [that] you can only say something about where people are right now and not necessarily about where they might be in their perception, their attitude, their acceptance a year from now because the technology is developing. As we go along, both what's good and bad consequences, people are reacting [and] looking back to technology now. So, it's true that having an accepting [mindset], I don't think that necessarily means acceptance of the highest level of technology right now...Before we only asked about self-driving cars and used the term self-driving cars because we thought

---

[25] Americans still don't trust self-driving cars, Reuters/Ipsos poll finds, April 2019
[26] Cox Automotive, Evolution of Mobility: Autonomous Vehicles, August 2018
[27] ORC International and Advocates for Highway and Auto Safety, CARAVAN Public Opinion Poll: Public to U.S. Senate: Pump the Brakes on Driverless Car Bill, July 2018
[28] Smith, A. and Anderson, M., Pew Research Center, Automation in Everyday Life, October 2017
[29] Advocates for Highway and Auto Safety, November 2019. Link here: https://saferoads.org/wp-content/uploads/2020/01/AV-Public-Opinion-Polls-7-22-19.pdf

> people... it would be the best way to communicate with survey respondents... And even in the worst survey that we did um we used the term self-driving car although we did, that was the first time we actually noted that personally owned cars versus ride hailing cars versus car sharing cars. So, we found significant differences in people's perception among the different... what we call the different applications. But I really liked and [was] very interested in those research [projects] for AAA because there, you are actually able to look at...the way in which people's perception of the different levels of technology whether they were consistent or inconsistent.

This approach contrasts with the "forecasting" mindset that is common in many discussions of AI's long-term development, in particular the possibility of a "superintelligent" AI system.[30] The narrative frame that determines how AVs are marketed to the public actively shapes the public's expectations of what AVs will be, which in turn reshapes the business models of manufacturers and startups, changing how and when certain technologies are incorporated into AVs and feeding back into the public's anticipation of their capabilities. The complexity of these dynamics make scalar prediction of public acceptance impossible, instead suggesting the need for a temporality of sustained public engagement:

> I was reading many papers and documents where people were just speculating on what was gonna happen in the future and I wanted to do something different. I wanted to start charting and monitoring how the public was reacting to the technology. I knew that we couldn't really say how the public would react or consumers would react because the technologies were not available, but we could start looking at and tracking those kinds of questions over time.

Moreover, public perception of lower autonomy capabilities serves both as an indicator of adoption of fully-autonomous vehicles and a guide for how to roll out autonomy capabilities differently to better meet or shape consumer demand. Even in a strictly technical sense, public surveys about adoption conditions are just as important to AV development as computer vision:

> Actually one of the reasons they're talking about self-driving vehicles, highly automated vehicles, is that the technology was so new that we felt it was the best way to communicate with the general public about the technology. But I think they're very interested in how the public is reacting in dealing with automated technology, that...level of automation which we promise... If we truly want to understand the impact of highly automated vehicles for being need to start again, tracking and monitoring how people are interacting with these lower levels, you know, technologies that Level

[30] Bostrom, Nick. "Superintelligence: Paths, dangers, strategies." (2014).

> SAE Level 2 and 3 which are actually out there in the market that people are actually using. We just finished a study where just a final part of the analysis a study towards AAA Projected Safety, actually looking at people's perception and use of technology 2 and 3 and their likely use of 4 and 5 and their perceptions of the technology where it relates to safety.

It is not possible to understand the technical development of AVs without tracking the parallel development of a public interest in them:

> My thinking is that it's possible that as the technology, as these vehicles are on the road and they're gaining more experience that they are learning because they have to... they learn by doing, that people will start to understand that and then, you know that distrust will go down, but right now it's up because of where we are in the development process.

And yet, automation levels are poorly or not at all understood by the public at large. At car dealerships, autonomous capabilities are typically advertised (e.g. lane and parking assistance) without reference to them. The levels are instead a means for different sets of experts (policymakers, academics, human factors researchers, tech enthusiasts) to talk about technologies as they become implementable and ready for market penetration. In other words, they comprise a narrative frame which the public cannot share or modify. Moreover, educating the public is not an option, as even the capabilities of present-day vehicles are poorly understood:

> I think we need public education now for the technology that is on the vehicles now. So, not for Level 5 because we don't know exactly what that will look like or when that will happen, but we have technology on the vehicles now that people aren't knowledgeable about.

NHTSA's classification of autonomy levels does, however, serve as a technical metric for experts to measure comfort towards particular AV capabilities in the wake of major vehicle accidents:

> [Public outreach is not] increasing knowledge through knowing more about what these technologies are. What it's actually doing is creating that very dynamic level of acceptance...So, I see actual trust going down and acceptance to the higher levels of automation decreasing as people are learning more about... today there are Tesla autopilot fatalities or the Uber fatalities that happened in Phoenix and because those are reported very widely and the negative consequences that are reported very quietly by the media. So... but it could be that, you know, that's where we are today. So for instance, AAA did a survey and at the end of a

> national survey at the end of 2017 and they asked the question of ‘would you be afraid in a totally self-driving vehicle’ and something like, 63% of the sample said, "Yes, they would" they repeated, replicated that survey in the summer of 2018 after the Uber accident and that 63% increased to 73%. So... and the largest increase was among millennials and distrust in 2017 it was only 9% would be afraid. In December of 2018, it increased to 64%. One of things that we do, we've done in most of our surveys when we just finished the one for AAA foundation for traffic safety, we might do a survey but then leaves [it at] the survey with telephone interviews of respondents, and in the AAA survey, we followed up with, we did a hundred telephone interviews so I think the samples are three or five hundred or something and those we did follow-up [with] like a hundred people where we probed on some of the things worth finding and it was the same thing, we're finding a lot of people who are, you know, their comfort level was more at Level 2 or Level 3. Not at Level 4 or 5 and they specifically mentioned media stories about the negative consequences of the accidents and the fatalities.

Surveys also serve to illuminate or falsify anticipated generational effects of AV adoption, suggesting new marketing strategies sensitive to a willingness to forego privacy protections and accepting new forms of data surveillance:

> We did... that survey... 2015 survey, it was a survey in Austin, Texas and the next year, we followed that with surveys in Dallas, Houston and Waco. So, we had surveys from 2016 and then in um 2018 is when we did the survey for Lyft and now, we interviewed in San Francisco, Phoenix, Las Vegas and Boston. So, we had data from different places in each we use the same model to help analyze the data, and what we find is if you just look at the descriptive, you do see that there are some differences, often not significant but there could be differences in acceptance by age, but when you control for other variables such as concerns about data privacy or um your attitudes towards technology um those differences go away. So, for instance we found out one of the most influential variables in being um and your likelihood to use self-driving vehicles is actually following the adoption codes. So, in early adopter of technology in general, will be an early adopter of automated vehicles um and there is a somewhat of an age bias to that. Younger people tend to more likely be early adopters but it's not their age, it's this propensity to be an early adopter that is this uh kind of factor and um we and that one of the reasons why we tested that with the Lyft survey because Lyft actually came to us after reading our findings in previous work and said, you know, ride hailing is a mobility technology, we wonder if early adopters of this ride hailing technology will be early adopters of self-driving vehicles and that's where we follow the very significant relationship.

This work apparently captures consumers' preference adjustments in response to real-time events. But there is a key indeterminacy in the difference between descriptions of AVs as presented in such surveys and the native concerns, assumptions, or perceptions of risk in the survey respondents. This indeterminacy is crystallized in how surveys treat AVs as a *marketing frame* into which respondents are free to project their own economic, cultural, and generational anxieties, even as surveyors learn how to better manipulate those projections through adjustments to the narrative frame.

This indeterminacy is readily apparent in the tension between perceptions of trust and comfort of AVs. While AV survey researchers often treat these as the same concept, their findings contain paradoxes about which AV capabilities are preferred vs. which are believed to be safer:

> A lot of the follow-up interview questions were in areas where we saw some inconsistency...that I'm still trying to figure out is, people were most comfortable at Level 2 or Level 3 technology but we ask because we did this for AAA as part of their traffic safety culture index and our job was to create questions on vehicle technologies to add to the traffic safety culture index. And so, in that index, in that survey, they had these driving situations like driving while texting, driving under the influence of alcohol, driving under the influence of drugs, speeding, you know they have all these driving safety conditions if you will. So while they were most comfortable with AV Levels 2 and 3, they also said that Level 5 vehicles would be the spaces in each of those driving conditions. Did that make sense? So, it's sort of a paradox. They find that, there'd be greater safety with Level 5 but they're not comfortable at Level 5, they're comfortable at Level 2 and 3.

These findings have prompted speculation about the grounds for these public sentiments. Limited follow-up work has appealed to problematic analogies made by laymen between AVs and other technologies with which they are more familiar. These analogies generate a complex moral calculus based on personal fear, a drive for individual control to mitigate that fear, and a general optimism about the future state of AV development:

> There are several elements of the technology... that makes people uncomfortable. One of the things that came up a lot is this thought that technology may not function. So, people equated it to their smartphones or computer that often has glitches and they saw the automation in the vehicle as being like a computer in your vehicle and, you know, my laptop has all these glitches, you know, [it's] feasible that my car could have glitches and I don't want that, you know. So, I want to make sure that I can take over the vehicle. So, I want to be able to override the computer

> that's in my car. So, there's this level of where they feel the technology is now which makes them the lower levels of technology but I think, they can see in the future if the technology is working as it should, you know, perfectly reliable that it will be safer. So, they have... they're optimistic that it will get there in the future, you know, have this very positive impact on safety but they don't feel that the technology is there yet so that's why their comfort is at the lower levels.

When jammed within the technocratic frames deployed by surveyors, these sentiments amount to a general sense of illegitimacy and anxiety about the future: "Even if you believe, in the end what would be the safest. 'I do not feel comfortable. I think it's good for the rest of society, but I do not feel comfortable.'"

In such a situation, safety concerns win out by default, not because they are more important but because they are easiest for experts to define and measure. Perhaps more importantly, they serve as a guide for how to nurture public trust in particular automated capabilities as they increase in scale and complexity. This in turn enables firms to distinguish between failures of technology and failures of public adoption:

> I think the largest component of the public interest is the safety component. I think although we don't yet have repeatable evidence that they will be safe...and we have less evidence for some of the other potential benefits whether they be mobility benefits or environmental benefits. We don't have evidence of those yet so in my mind those could still go either way. But it's one of the reasons why I changed, number one we really need to enable these vehicles to be on the road as test vehicles. They need to be tested as much as possible. That's really important and that [for] public interest to do that it's how they're gonna get evidence. And number two, I think it's ... that's why it's really important for us to be looking at what's happening at the lower levels. So, for instance, you know, where we have Level 2 vehicles on the road with parking assistance, collision warning, or forward collision of... are these technologies actually bringing safety benefits, if they are that's great, if they're not, **is it because of the technologies themselves or is it because people aren't using the technology**. I have the technology turned off whether or not purchasing the vehicles with technology in them and the same thing could be said for Level 3. You have to test the autopilot [or] Super Cruise but people could purchase like right now. Are those vehicles that are actually operating on the roads right now having the public benefits that were promised? Whether they'd be safety or mobility, or you know, traffic flow, those benefits are really being seen, and I don't think enough people are really looking at that. (emphasis added)

This approach leaves the survey community in the somewhat reactionary and passive position of identifying why the public remains skeptical of particular AV capabilities, and how to better inform them of the demonstrable public benefits (e.g. safety) they offer. What surveyors are not doing is working to identify what vision of transportation or urban living the public wants to be enacted by the capabilities in question. Such a vision needs not be spearheaded by the federal government, and may well amount to informing the public more actively about local transportation as it already exists. Cities could then more actively consider the infrastructure options available to them as AVs expand the infrastructural possibilities of transportation:

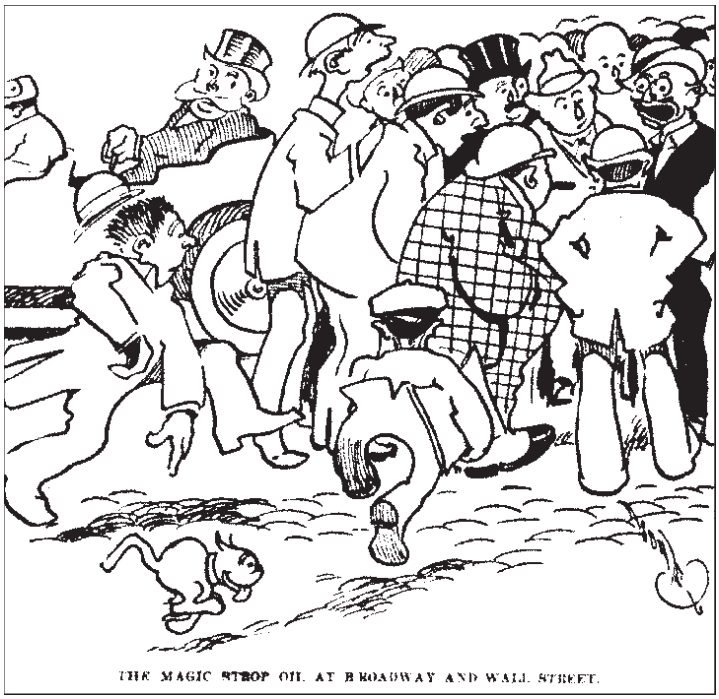
> I see it as almost like a public awareness campaign like, you know, don't litter. The anti-littering campaign or any kind [of] public education campaign that would... but that would be run at the local level. So, for instance, you know, we have these pilots going on in different cities that would soon like this city, would want to educate the public about the pilot vehicles that are learning on that street. I'm here in Austin right now, we have a low-speed shuttle running now [inaudible] um be careful about jaywalking from the vehicle. You know, so I see that as a local initiative or the pilots we're running... but I see the need for someone, be it the auto-makers or the dealerships, someone from that side educating people as to the technology that's on their cars now.

Such an approach could be augmented by real-world testing of AVs, rather than relying on computer simulations to validate software. In this vision, unknown unknowns are handled by educating the public, not by altering the model specification or leaning on data:

> Interviewee: I think for the importance of real-world testing is... in that the unknowns that you encounter when you're in the real-world as opposed to any kind of simulated environment. So, for instance... I mean, I think the testing on roads is not only the learning of the vehicle in that environment but also the mapping of that environment. So, it's sort of serves two purposes and you know, we know that the details, digital maps are really important for the vehicles to operate safely and furthermore driving that they do on the roads or even in corridors the more details those maps... that the vehicle can rely on become and I don't think you can do that in a simulated environment. I could be wrong but I don't see that happening and there is this unknown factor I mean I guess you can try and come up with any number of scenarios and drill it the vehicle in the simulated environment but I guess there would still be things that would come up in a real-world environment what wouldn't happen [in simulation].

> Interviewer: Would it be fair to say that your intuition there also resonates with the kind of much greater focus and trust that you have on educating the public in the short term about these vehicles, precisely because there's so many unknowns?
>
> Interviewee: Right, exactly.

The politics are likely to manifest in the *jayification* through which AVs will be understood or accepted by different publics, as private companies, consulting firms, and municipal entities can encode descriptors into surveys as they see fit and thus shape the types of consumer demand that suit their own preferred business model and form of public-private coordination. This would mean that different consumer brackets are given structural incentives to self-select out of using public roads, and likely to think of themselves as a 21st century equivalent of jaywalkers if they ever do find themselves making use of them.



The jayification of public roads, over 100 years ago.

Consumer surveys will continue to reflect some narrated understanding of how prospective consumer purchasing habits map onto various models of AV performance. Regardless of how accurate or well-motivated this narrative is, it structurally compromises public mobility for the sake of marketing purposes. This reflects a truncated form of communication and engagement that mobilizes the public only to the extent that it can be modeled as a reliable consumer of a private good rather than an active participant in defining the future of mobility.

Indeterminacy 3 - Selective robustness to signage

Computer vision, aided by sensors such as LiDAR, is widely used in order to meet specified metrics for image classification, including those related to object detection and collision avoidance. In tandem with cameras (on which Tesla Autopilot exclusively relies), radar, and extensive mapping of road signage, these sensors aid in what roboticists call localization: the process of determining where a mobile robot is located with respect to its environment.[31] Intuitively, the aim of this agenda is to help AVs accurately perceive and respond to external environment features as an approximation of optimal driving behavior.

However, it is not always clear how or whether the chosen evaluation metric maps onto the chosen specification of what it means to be a good driver. This is for the simple reason that an automated vehicle--unlike a novice human driver, or even a child--has no understanding of why external signage exists. The ability to reliably predict what other road users might do, no matter how refined the assumptions or rich the dataset on which those assumptions were validated, does not equate with a counterfactual grasp of how else the world has worked or could be made to work. Consequently, the best behavioral model in the world could still not be trusted to respond appropriately when key features of the driving environment (on which it has previously relied) are compromised or absent.

This problem is crystallized in the research agenda of *adversarial learning*, particularly as it comes into tension with with more structured and specified vision architectures for AVs. In brief, adversarial learning is an optimization procedure that mobilizes "malicious" examples and strategies to derive a model that is robust to the most significant variations in the underlying environment features. The model then serves as an approximation of the "true mean", rather than just naively minimizing some cost function in the absence of a trustworthy data distribution. In other words, it is a way to train AVs to more reliably recognize road signage by learning how not to be fooled by subtle deviations (e.g. dirty stop signs are not misclassified as something other than stop signs). Put even more simply, adversarial learning is a way to train automated vehicles, despite not knowing why stop signs exist, to get better at seeing them by showing them lots of images of things that look like stop signs but aren't, and confirming that they know the difference. This is different from training them on all the images of stop signs we have, as the latter may not be backed up by strong assumptions about the data inputs.

The context behind this intuition is explained in more detail by the rise of deep learning optimization techniques, which supplanted the more "classical" design of algorithms that had more explicit assumptions, hard-coded rules, and constrained specifications:

[31] See for example Montemerlo, Michael, and Sebastian Thrun. *FastSLAM: A scalable method for the simultaneous localization and mapping problem in robotics*. Vol. 27. Springer, 2007.

> Let's say you want to process images, and let's say you want to reconstruct the objects you're seeing if you're in the self-driving car. And so you see like, "oh, there are some people there, there are some buildings there, there are some signs there". I would say that the modern Deep Learning approach to this is that we'll use a very holistic architecture where you take this image and then you do some kind of a...whatever convolution they're calling and then they do some kind of max-pooling and then and then in the end they get a pixel by pixel description say, "oh, this is a person, this is a tree, this is something else". Whereas, the traditional approach would try to be more hands on about this, they will try to recognize what kind of feature or what kind of distinctive characteristic a person will have, and they would use this kind of human observable thing to characterize what is in the scene. And there are benefits to both of these [0:04:00] approaches I would say...The traditional approach is more like defining special cases over algorithms like if you have these features, do this, if you have these features, do that. While in Deep Learning, you don't care about these special cases, you take a very holistic approach.

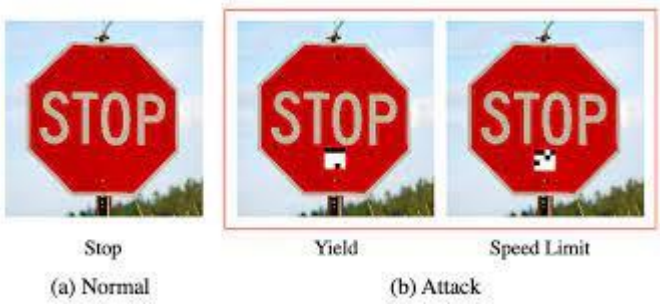
It is hard to overstate the technical success and influence of deep learning as a potential general paradigm for image classification. Yet machine learning practitioners are not in the dark about its limitations, for example with respect to its qualitative differences from human perception:

> If I add some random noise in the figure, then my eye cannot detect everything, detect it because it just looks the same. For example, I have some points here. Like you still, you can still tell it's a laptop but for machines, because they're trained like that, so it doesn't know what he means if you add some, like, random noise because that would make a machine, because the machine would tell us one as totally different from the previous one. So, then the machine is really hard to tell between the perturbed image and the original image. So, my point is more like, because the machine is totally trained differently from the car and the human. So, we cannot expect the machine to calibrate with humans, even though they're [comparable in] performance. There are some claims that say ‘the machine is really beating humans in terms of computer vision and natural language processing’, but that's only one-dimensional projection of their performance. You see, you cannot just evaluate a person according to his height or according to his weight or something. Because everything is like multi-dimensional things, it's also your evaluation score.

Deep learning also as a reputation for being unprincipled and excessively data-driven, sparking efforts to combine it with more traditional approaches:

> What you would do is you would collect more data related to that person and then tune your architecture to achieve what you want. But that is kinda like [a] whack-a-mole thing where you fix this and then you will find other cases that you were doing well that you now don't do very well, and so debugging is a serious issue in modern Deep Learning approaches. But using these special cases approaches in the more traditional methods, I would say that you are able to recognize these problematic places that your architecture cannot handle well, and deal with them specifically. So in the long-term approach, we want this in something like a human brain with a general Artificial Intelligence, but I would say that [general AI] and short and mid-term having a very long list of special cases is a more immediate way to solve all the problems of debugging in Deep Learning. So our research [is] trying to recognize what is special in Deep Learning and we figured it's that you're always computing gradients of some models, so why don't we just generalize the model and also take the traditional methods as one of the models.

Intuitively, adversarial learning is an attempt to statistically compensate for the differences between computer and human vision. It attempts to push the limits of learned models on the assumption that any such model is a confused approximation of the real world, as it is constructed from statistical learning based on scalar values rather than the "higher dimensional" nature of human perception. A machine, for example, can only compare relative errors of classification, and cannot communicate confidence in its classifications *as such*.



Adversarial learning applied to stop signs. Here, minor adjustments (which a human may well ignore) cause a neural net to misclassify the image.

> Because [artificial and human intelligence] are just like two, totally different things and you're trying to say, so okay, so for now, they're doing the same thing on classification, one task. So, they're doing the same thing in classification, like the Artificial Intelligence is able to tell you and human is also able to tell this [is a] panda. But, in fact their internal rule is

> totally different. So, humans are more like I invest some knowledge and then I just try to connect that with my previous prior and try to tell things like this and...intelligence is like, I learn from data that there are some features of the image and then I tell from the image that this shall be a panda. I tell from that feature that this image shall be a panda. That we are just trying to only compare their classification result procedure. Because the procedure is more deep in that, it's really hard. Human intelligence gets evolved like, generated from scratch. But then, Artificial Intelligence, they just skip from the whole procedure. They just say, like I give you some data and you give me some result. Then [but] for this it's really hard for [it] to be really robust in terms of human perspective.

For these reasons, it is extremely difficult to train AI models to well-approximate the situational sensitivity and complexity of human vision. The paradigm of adversarial learning responds to this by empowering the designer to pit one AI model against another that is trying to fool it. The first model's problem is then to ask itself: "what is the optimal strategy I can do to prevent you from screwing me up too much?" This allows the designer to squeeze a lot more value out of a dataset whose distribution of images is already well-understood and accounted for, effectively permitting the model to validate itself through a competitive form of marginal self-improvement.

However, the benefits of adversarial learning are asymptotic, as there will always be more examples to challenge the model optimization absent a determinate conception of the environment in which the task is modeled:

> Adversarial robusticity is a very, very, difficult thing to achieve, because if you fix this [set of] adversarial examples, people will just find other adverse examples. [There is] tons of research that does this defense and attack thing and so far, the defense has never won, and so I don't want to sound pessimistic but I feel that in the short term, we don't see a very clear way of how to just solve this by generating more adversarial examples. We need fundamental changes to the model to fix all the cases.

Moreover, adversarial learning is known to scale more easily and appropriately for image recognition (e.g. classifying observed objects as stop signs) than for motion planning, where there is less tolerance for error "because as long as you make some mistakes, then you just kill somebody". Concretely, this means adversarial learning is much more tractable for recognizing static images of stop signs than for evaluating different paths an automated vehicle could take through an intersection. In general, an extremely strong prior is needed in order for adversarial learning to succeed, as the underlying source of the data distribution must be trusted in order for

errors to be reduced in a way that is known to map onto reasonable assumptions about the real world. Without that prior, adversarial learning is little more than a hammer in search of a nail:

> Motion planning is more from a control perspective. You have something that you already know very well, and you believe everything shall be correct. That's a clear assumption you have. It's like, your map shall be correct and your information about other cars shall also be correct, but given that assumption, which I believe is kind of easier to achieve in image recognition. Then you can just say, 'oh okay, I would just go this way and go this way', something like that. Then the only thing you need to deal with is, there are [for example] some people who are walking in front of you. That's the hardest task. I believe that's already a very much easier task than trying to achieve everything by image recognition.

This means there are two major limits to the adversarial learning paradigm. One is the problem of reality itself, in defining the interface between the learnable model and the system(s) humans themselves use to interact with the world. Second is efficiency of computation with respect to modeling precision, which makes it at present unworkable for motion planning:

> Just because you perturb the images doesn't mean that this image will actually appear in real life. So, there's one possibility that, first of all I don't know why a human can still recognize it, it's kind of mysterious to me. [Also] there's no guarantee that the space you want to cover with adverse examples [is not] much larger than you actually need to cover, and that might turn out to be impossible. [This is also] because the space is so large and their limited computation resources. Second of all, the power of gradients has shown to be able to attack any kind of defense people are applying. So you can always find new adversary so this is kind of related to first one, that the space is just too large that no matter what fix you do, you're just not able to, you probably will not just be able to fix all the possible cases that appear. So, my feeling is that people are trying to solve a harder problem than they necessarily need to solve.

> Deep Learning is too slow for some interacting applications, but a lot of the models still use traditional computation. Sometimes it's cheaper and sometimes it's more expensive. So, you have this efficiency constraint. There are a lot of very useful and powerful techniques you can't use because your application requires a very restrictive time constraint, and again self-driving cars are one of them. If you want to do some sort of inverse rendering to reconstruct the whole 3D scene and you're driving a car. It might be possible but, at the moment I don't see how it's practical, so you probably can only reconstruct the very core street view of your

> scene. You probably don't want to reconstruct every leaf of a tree, so you probably want to process the tree as a cylinder or something. So there's that, and we need to figure out how to approximate this model in order to make it able to run under the real-time constraint on some systems, and that part is not solved yet either. I think that's also a problem with Deep Learning approaches.

It is worth clarifying that, as a historical matter, rules of the road were adopted to aid in multi-modal traffic coordination, not to make individual driving behavior robust to perturbations. By extension, the purpose of signage is to support and corroborate public traffic conventions that are known to and accepted by all agents, not to literally populate roads with images that--if they are sufficiently well-recognized--serve as a meaningful proxy of good driving. But the limits of adversarial learning suggest AVs are likely to be made robust to *particular forms* of road behavior and signage that are easier to model or whose data distribution is better understood, regardless of their normative content. This may well leave other forms of mobility that are dependent on other features (e.g. pedestrians) vulnerable to the AV specification, perhaps generating unfair coordination conventions.

This means that the prohibitive computational expenses and liability issues of deep learning for particular driving tasks introduces a distinctive mode of sociotechnical politics. Consider the example of designers justifying their choice to make specific layers of the neural net informed by the physics of the domain rather than computational efficiency. This appeal to physical intuition, however well-motivated, ignores the normative content of roads, as apparent in the pseudo-alchemical justifications of practitioners:

> It's a kind of jargon that usually you can take an image and then you would and then you can process it either using the convolution on your network or something else, and then you would arrive after the intermediate layer. And the question is, what do you store here. So a lot of convolution on your network to store like, a bunch of numbers that are not super meaningful, but I would argue that--I'm not the only one that's doing this, but I'll argue that you want to make this intermediate layer a kind of physically inspired layer. So, what you will do is you would take an image and then [0:27:00] this is just one way to achieve what you're saying, and this is not like the only way. I get that this is the short... probably the more promising way. And so you would take an image, and then you have these intermediate representations, and instead of numbers you would make them like, physically meaningful. Like, you would say that ‘oh there is a tree here.’ ‘There is a pedestrian here.’ ‘There is a building here.” And then you can also model human motion. There are statistical models of human motion which you can gather from other places, and so this can be part of the intermediate

> implementation as well. You can say, ‘these people were coming from Denver so, it's probably going in this direction if time goes on.’

To illustrate this politics, it is necessary to appreciate the vagueness of *robustness* itself. Defining the ability to withstand adverse conditions first requires active definition of the task being performed and how to model it well, even prior to the selection of metrics for optimization:

> [O]ur work is more like, we already know what is the optimal error you can achieve in this. And you try to achieve this in the easier task, but for harder tasks like driving, it's really hard for you to give the optimal thing then it's hard for you to define what kind of, what level of robustness you are satisfied with. Because for this task, you know the best you can do. But for real-world applications, it's really hard to give [guarantees]. **So, then with robustness we are more about our goal of defining the question.** If you really want automated cars to be used in practice and widely, then robustness, I guess really means you need to be perfect in terms of making all the decisions. Yeah, it's very hard even for me. I cannot believe it will happen, but I guess, at least it means much more about exceeding human performance. It should exceed human performance much, much better than we can say like, yeah, you're good enough, you can just go to the real world. Otherwise, I guess it's really hard for automated cars to survive if you're just achieving the same performance as humans, but humans would not believe that. Because there's also another reason. Even if you're optimizing your error rate to be as small as possible, you're still just evaluating it under a very small subset of the real scenario. It's really hard for other guys to trust that you're really just making everything the same. [...] So, if you want an automated car to drive well in different areas, then robustness also means you need to adapt to different areas, situations like driving style or something. Humans can easily adapt to that. But I guess for a trained system, if you only fine-tuned on that, I guess it might be hard because...they have very fixed parameters for everything. That's also a way, so basically robustness means, you need to behave real under all of the extreme situations that it is possible in simple tasks. But in harder tasks, we cannot give a theoretical guarantee. Now, we have to resolve to experiments. But experiments for now are only very limited in terms of this situation, in terms of the number of experiments you conducted. So, it's really hard for you to convince people by experiments. That's an issue. And also, if you're really brave enough to give some real-world test and if you really screw it up, then you're just, your company would be just dead. People cannot give you a second chance of trying. (emphasis added)

The specific source of indeterminacy is the effort to define, through model optimization, a good translation of the road environment. No model--even those made possible by deep learning--is

capacious enough to capture the world state as humans experience it without reinterpreting it in the eyes of the designer:

> I would say that the biggest challenge of all this is, what is the correct approximation. You introduce all these physical systems in order to [find out] where you can do it computationally. Where in this combination will be feasible and where [is it] accurate enough for real-life usage? And this is again, I don't know what to think. So, if you look at the simulation people do in those self-driving car companies and you look at like, Pixar movies or Disney movies. The quality is very different, and why? Because one frame in the Pixar movie takes like 8 to even 24 hours to render, it's just one frame and there are like 24 frames per second, and it's a 1-hour movie. For that, they have to [prepare] for 1-hour portable simulation, they have to render for maybe 1-year CPU core hour. 1 year of CPU core computation. So, and that's just 1 hour that probably can't cover too much of training data or verification and so, we can't do very, very detailed and high-quality simulation. And you look at like, [IA] like they can do everything in like 60 frames-per-second. It only works in that specific environment, and when you're trying to make this work in all sorts of situations, then a lot of human engineering is so hard. And [it is] all human engineering to find the right approximation, is required in those [situations]. So, they have these artists trying to fade effects, they have programmers trying to speed up crop hedges in certain cases to fake stuff. Like if the tree is too far away, just render it as a cube or something. So this kind of stuff is necessary, and it's important for any kind of old problematic testing or simulation on any kind of system.

The technical intractability of robustness exacerbates a public problem that is already well-known to traffic planners: *damage to infrastructure*, in particular road wear. It is the product of a development situation in which the practical achievement of preferred optimization metrics implicitly prioritizes certain forms of signage, movement trajectories, or patterns of traffic engagement for which formal guarantees are easier to give, or at least approximate. Over time, this will naturally condition the integrity of the road environment on the operation of the AV fleet, rather than the other way around. And at some point, it will become easier to redesign particular roadways to be suitable for AVs rather than AVs for public roads.

Here the emergent form of sociotechnical politics is what we call *rubblization*.[32] Traditionally, rubblization is an industrial process for building new roads by reducing existing roadways to rubble and then laying them over with asphalt, rather than removing older road segments and replacing them with new ones from scratch. This has the effect of saving money, material resources, time, and energy emissions. Rubblization is a means of translating and remaking

[32] https://en.wikipedia.org/wiki/Rubblization

aspects of the existing road environment into a new form that befits the needs of the present moment, and saving costs while doing so.

Rubblization in action: pummeling concrete into a base for new roadways.

Rubblization, as a sociotechnical process, reflects how AI systems interpret pre-existing physical infrastructure and institutional norms as a resource to be observed, processed, and made use of in the interest of optimizing task completion. The "task" in this case is, by default, the regular and safe completion of road trips. Any unusual road situations, objects, or traffic dynamics that stand in the way of that task are likely to be "rubblified" as AV fleets are rolled out. This may take many forms. For example, potholes and other types of physical road wear could arise from a given AV fleet that learns to navigate particular streets at greater frequencies than others. It could also mean that unusual streets--referring either to signage, turns, inhabitants, or the types of interactions that an AV could be forced to enter--become problems whose resolution is indeterminate without additional criteria from the public to specify what the integrity of roads actually means.

AV developers have a few choices about what to do in these situations. The most obvious is to ignore the problem and optimize the cars' performance for other streets and signs that are easier to model. In that case, the path forward is to neglect basic infrastructure as needed. Another is to treat these situations as challenging "edge cases" for vehicle fleets, grist for the optimization mill. In that case, it seems likely that wealthier and more predictable neighborhoods or city districts will be the first (and perhaps only ones) to reap the benefits of AV service, as other regions will be too expensive or poorly-modeled to reliably test. A third is to treat them as problems that must be solved in order for AV fleet deployment to proceed legitimately and purposefully, resulting in their reappropriation and refurbishment as part of the built

environment. This would likely mean that AV firms and startups would partner with municipalities to reform these streets to be navigable by AVs in some form.

Rubblization refers narrowly to the last of these choices, and it is an aspirational mode of politics. Rather than communicating to the public the inevitability of AVs, or persuading the public to accept AV testing as the price of being part of the future, rubblization could comprise an example of Grunig's two-way model of public engagement: public and private stakeholders communicating to identify mutually beneficial terms on which AV fleets could be integrated to neighborhoods. In fact, one deep learning practitioner endorsed a similar strategy for AV companies to follow, rooted in developing methods for uncertainty estimation in safety-critical settings that analogize from how human drivers deal with their own limited perception of physical objects:

> A lot of people are saying that [we need] some sort of uncertainty estimation of your model. So, as a human you can see that, because all this image is like all white, it must be over-exposed, and so having these kind of things in your system, like having an uncertainty meter if you see that the image is super over-exposed, you can probably expect you are going to make wrong decision later. So, having this and doing the most conservative action you can do at that moment, in that case you should just probably just brake because you can't see anything and that's what a human would do if he can't see anything. Having some sort of uncertainty estimation is, I would say, necessary for a lot of the safety-critical applications... And if the observation deviates from actual physics a lot, then you should be suspicious that your observation might be wrong. So, physics can add a strong primer for uncertainty estimation.

This suggests a possible role for the public to play: defining uncertainty estimations in indeterminate traffic situations, once safety-critical features are ready to be optimized via adversarial learning.

Discussion

HCAI is able to provide a throughline from AV development choices to current and emerging forms of sociotechnical politics, while also implying what more responsible and stakeholder-sensitive design commitments would look like. It is beyond the scope of this paper to trace out an exhaustive set of governance mechanisms that could corroborate or enact these commitments, let alone endorse a particular implementation strategy. Rather, in this section we have highlighted a) the emergent sources of indeterminacy at stake in different stages of AV development; b) the public problem at stake in each of these stages, reflected in particular design or framing choices; c) the corresponding group of technical experts responsible for representing the public interest. Throughout, we have highlighted the paradox that this interest cannot be represented until the

public has been organized, and the public itself cannot be organized until it has made sense of its own vulnerabilities by evaluating the relative behavioral consequences of alternative AV specifications. While we have sketched some of these potential consequences under the labels *moral crumple zoning*, *rubblization*, and *jayification*, it is ultimately up to the public to anticipate and decide which consequences follow from available choices, how to evaluate those choices, and which are worth pursuing.

Towards a New Theory of AVs and the Public

The insights explored above from the interviews and analysis do not exhaustively cover all indeterminacies found in the development and integration of AVs. However, they do form sufficient evidence to motivate the formulation of theoretical lenses on what the public problems of AVs amount to, and what it may take to inform and organize the public itself to work towards those models of the future of transportation that are its to envision.

To do so, we mobilize the philosophical methodology and political theory of John Dewey, most famously captured in *The Public and its Problems,*[33] in which he sought to refine inherited definitions of the "public" and the "state" in the context of social and technological transformation. Dewey's deep sympathy for democratic society and governance, which we share, had both analytical and normative dimensions, and we mobilize both in order to apply his insights to the AV context.

We first cover Dewey's key notions of a public and public problems. We then interpret Dewey's insights for the emerging public problems around AVs. The section ends with reference to the situational theory of publics, proposed by James Grunig. His models of public relations show how to apply our HCAI framework to the politics of AVs, suggesting forms of feedback to empower rather than subvert the public's efforts to make sense of its own problems on its own terms. Together with the analysis in the previous section, these theoretical insights will help draw implications and recommendations for empowering the public to make sense and adopt a more anticipatory stance towards the possibilities and public problems that AVs bring into existence.

Dewey's Public and its Problems

In his seminal work *The Public and its Problems,* John Dewey summarizes the public's state of being in the following passage:

> Each form of association has its own peculiar quality and value, and no person in his senses confuses one with another. The characteristic of the public as a state springs from the fact that all modes of associated behavior may have extensive

[33] Dewey, J. "The Public and its Problems (Athens, OH, Ohio University Press)." (1927).

> and enduring consequences which involve others beyond those directly engaged in them. When these consequences are in turn realized in thought and sentiment, recognition of them reacts to **remake the conditions out of which they arose**. Consequences have to be taken care of, looked out for. This supervision and regulation cannot be effected by the primary groupings themselves. For the essence of the consequences which call a public into being is the fact that they **expand beyond those directly engaged in producing them**. Consequently special agencies and measures must be formed if they are to be attended to; or else some existing group must take on new functions. The obvious external mark of the organization of a public or of a state is thus the existence of officials. Government is not the state, for that includes the public as well as the rulers charged with special duties and powers. The public, however, is organized in and through those officers who act in behalf of its interests.[34] (emphasis added)

From Dewey's presentation, we extract a theory of the public as having the following mature features:

-the public is an agent. This means that the public acts on its own terms as a form of association rooted in the perception and recognition of consequences, and is something other than an amalgamation of individual preferences. One person can agree with a public's deliberations or share in its approach to problems, but to belong to a public is to share in its state of being and to have a stake in its problems. The public is therefore defined not by a kind of representation but by its potential to act as a single collective entity.

-the public has a particular interest. This means that a particular problem has emerged to be dealt with and confronted by the public qua public. This interest could very well be represented by a concrete person or institutional entity, but representing the public is secondary to acting as a member of the public itself. The latter requires that one share in and take on the interest of the public as one's own.

-the public, once organized, is the state. This means that the public can realize and affirm itself by naming certain interests as its own, reflecting on these interests on its own terms, and acting on that interest by establishing how various discrete tasks relate to the entirety of the activity. It is not a mere mass of people that in principle could act as a public, nor a person that acts as a magistrate on the public's behalf, though such officials are the "obvious external mark" of a public once organized. This is similar to how we say that a person makes or consumes dinner, whereas a brain merely follows a recipe, a hand lifts silverware, a mouth chews food, and a stomach dissolves it. The public's existence must actively direct and orchestrate the tasks carried out in its name.

---

[34] Dewey, John. *The Public and its Problems*. Pp. 27-28.

Where exactly do publics come from? Dewey argues the public emerges in an effort to construct "manageable limits" on consequential actions, stabilizing agreements between persons in order to predict and generalize the implications of conjoint behaviors:

> No one can take into account all the consequences of the acts he performs. It is a matter of necessity for him, as a rule, to limit his attention and foresight to matters which, as we say, are distinctively his own business. Any one who looked too far abroad with regard to the outcome of what he is proposing to do would, if there were no general rules in existence, soon be lost in a hopelessly complicated muddle of considerations. The man of most generous outlook has to draw the line somewhere, and he is forced to draw it in whatever concerns those closely associated with himself. In the absence of some objective regulation, effects upon them are all he can be sure of in any reasonable degree. Much of what is called selfishness is but the outcome of limitation of observation and imagination. **Hence when consequences concern a large number, a number so mediately involved that a person cannot readily prefigure how they are to be affected, that number is constituted a public which intervenes.** It is not merely that the combined observations of a number cover more ground than those of a single person. It is rather that the public itself, being unable to forecast and estimate all consequences, establishes certain dikes and channels so that actions are confined within prescribed limits, and insofar have moderately predictable consequences.[35] (emphasis added)

Dewey furthermore argues that the rule of law, rather than a primordial source of moral authority or political legitimacy in its own right, is in fact a "structure[] which canalize[s] action" taken by the public to define its own indeterminate situation. The public is thus not the object of a general interest calculated by elites. It is a mediate force responsible for both articulating the grounds on which general interests are representable as well as motivating the creation of contracts and specifications that establish the trustworthiness of agreements and determinate resolution of disputes. This in turn allows us to extract the key features of a public problem:

-the problem is specific to the numerical byproduct of vague agreements. This means that the public anticipates and stands for the trajectory of the common good. The public need not be tied to a pre-existing community, ethnicity, or culture. Nor is it restricted to a particular mode of communication, language, or set of signifiers. Rather the public is defined by a shared state of vulnerability whose scale serves to condition the existence of a given numerical entity. As a fact of the matter, the public exists whether or not its members are aware of the problem that coagulates their interest, because it exists as a state of consequence rather than directed will.

---

[35] Ibid., pp. 52-53.

-the problem's significance is determined by the actions of the public. This means that the public is able to mark the scale on which it exists according to terms whose range and intervals assign significance to, demarcate, and establish control over salient forms of risk. The public decides what is of value and what its values amount to with respect to the state of vulnerability that is its own. It analytically follows from this that the choice of utility calculus and prioritization of risks are the public's to make and to name; any approach to optimization or utility assignment assumes that this more basic allocation of value has already taken place.

Finally, Dewey notes that the organization of the public as an active political agent is a difficult process, made more so by the disruptive effects of technology:

> Progress is not steady and continuous. Retrogression is as periodic as advance. Industry and inventions in technology, for example, create means which alter the modes of associated behavior and which radically change the quantity, character and place of impact of their indirect consequences. [] These changes are extrinsic to political forms which, once established, persist of their own momentum. The new public which is generated remains long inchoate, unorganized, because it cannot use inherited political agencies. The latter, if elaborate and well institutionalized, obstruct the organization of the new public. They prevent that development of new forms of the state which might grow up rapidly were social life more fluid, less precipitated into set political and legal molds. **To form itself, the public has to break existing political forms.** This is hard to do because these forms are themselves the regular means of instituting change. The public which generated political forms is passing away, but the power and lust of possession remains in the hands of the officers and agencies which the dying public instituted.[36] (emphasis added)

The emergence of a radically new technology subverts the emergence of the public on two fronts. The first is that it alters the conditions under which people are associated and thereby the consequences that motivate the new public's generation. The position, magnitude, and scale of these consequences are unprecedented, yet must be discerned before the public interest can be deliberated upon and articulated, let alone represented by state officials. The second is that existing political agencies, themselves the residue of prior public concerns, will try to co-opt the new public's interest and conflate it with their own. Tragically, the more institutionalized and sedimented these agencies are, the more difficult it will be for the public to appraise its own situation and articulate what it wants as distinct from what the state already knows how to measure, represent, and value. Instead, the public must somehow organize itself by coming to terms with the effects of new technologies, and then reorganize and appropriate the existing state

[36] Ibid., pp. 30-31.

apparatus so that its agencies and technical capabilities (law, statutes, regulations, etc.) are made to work on its behalf.

AVs as a Public Problem[37]

Unfortunately, the present state of AV development and prospective regulation closely follows the faltering spirit of Dewey's claims. Let us review the present story of AV rollouts. We have been told that AVs are already here, that it is only a matter of time until they are widely adopted, and (notwithstanding some uncertainty) that this will happen soon. We have also been told that this is by default a good thing, as AVs are safer and more predictable than human drivers--after all, they do not get tired, or distracted, or inebriated. We therefore have a moral obligation to ensure that their adoption is widespread and smooth, chiefly by accepting their inevitability and by being willing to ride in them as soon as they are locally available.

However, each element of this story does not conform to the facts. AVs, understood as vehicles that can fully drive and navigate themselves on public roads, not only do not exist but comprise a disjuncture rather than extension of existing technical paradigms. This is because human designers, manufacturers, and operators must define and oversee the features of the road environment that AVs must attend to. In addition, AVs are not and cannot be safer or more predictable than human drivers, because they are not comparable to them unless their conditions of action and operation are further specified. To claim this is like claiming that the introduction of cars one hundred years ago made roads safer and more predictable because unlike horses, cars do not buck, neigh, or get scared.

What we are witnessing with the marketing and development of AVs is the state's consent to the corporate appropriation, definition, and attempted resolution of public problems. Private actors have persuaded a critical mass of federal and state regulators that their models will do a better job of representing and managing public problems than the regulators themselves. These regulators then agree to hand off what were once the critical functions of the state to the apparently augmented capacities of the private sector.

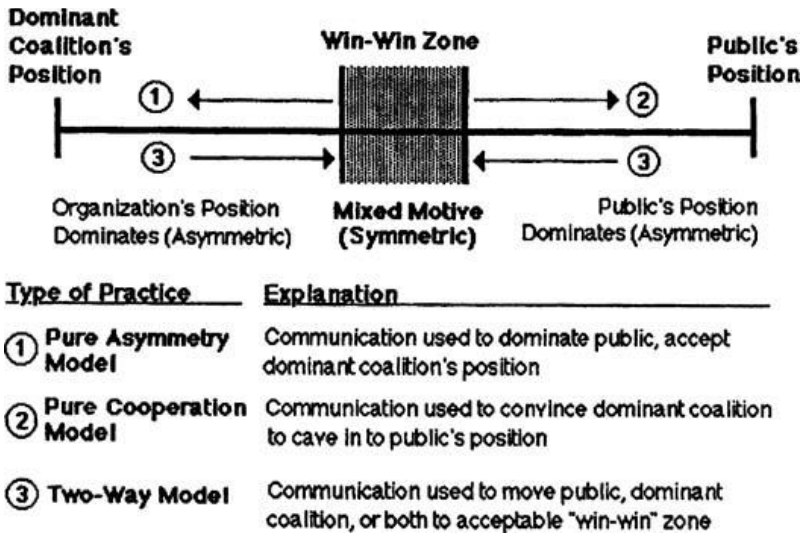
Dewey would have no quarrel with this handoff as such, because it is possible for public problems to change or be minimized over time in response to new facts and developments. If the private sector is able to step up and provide a service that was once the state's to perform, that does not mean the state has been eroded, but that the provision of service has ceased to be recognized as a public problem. The promise of AVs is to remake safety, the major flashpoint of vehicle regulation in the mid twentieth century, from an urgent public problem into a criterion of technical proficiency. But there have been other problems that define the automation of public

[37] This section remains in a more essayistic form. In a subsequent draft, we will bring it into closer conversation and interaction with the HCAI framework as well as the interview research.

transportation that matter here, such as infrastructural integrity, mobility, and liability, that are brought into scope through the reincorporation (or corporatization) of public space. The definition of what public space amounts to will remain a problem, and AVs bring more of the terms of that problem into scope for public deliberation even as they promise to resolve others. Hence AVs neither solve nor privatize public problems; they reframe them on new ground, providing the means for a new generation of public actors to canalize them.

Grunig's Situational Theory of Publics

James Grunig's models of public relations suggest strategies to apply HCAI to the politics of AVs, incorporating forms of feedback to empower rather than subvert the public's efforts to make sense of its own problems on its own terms. Building on Dewey's work, Grunig developed the "situational theory of publics" to account for how people can come to identify as members of a public based on their awareness of and ability to respond to situation-specific problems. Methods of communication can then be adopted to either inform people about the problem or about actions they can take to rectify or mitigate its effects. Key to Grunig's models is the intuition that different modes of audience communication organize a prospective public differently, serving either to manipulate, influence, persuade, or negotiate with whomever a given agency seeks to communicate. The types of feedback presented and proposed in HCAI conform most closely to Grunig's *two-way symmetrical model* of communication, in which conflicts are actively resolved with the aim of promoting mutual understanding and respect between the organization and pertinent stakeholders.



Grunig's mature model of public communication, identifying the win-win zone of mutual engagement.

We are particularly interested in Grunig's work as a possible model for AV experts to engage distinct AV-adjacent publics at various stages of maturation, including before they have had a

chance to articulate their own interests. Under what conditions can information be distributed and knowledge shared in order to encourage the public to intervene and name its vulnerabilities? Which agreements pertaining to liability, mobility, and infrastructure remain problematically indeterminate? Which publics are inchoate? Which are ready to be organized? In summary, how can experts begin to *canalize*, rather than represent or purport to solve, the public problems made tractable by AV development?

Implications and Recommendations[38]

Our appeal to the public is not rooted in moralizing about the need to take disruptive technologies out of private hands and place them in those whose hands are more deserving. While work to implement community-based approaches to AI system design is ongoing, laypeople are at present in no position to apply convolution to a neural net, employ investigative social science to measure their own preferences, or design research studies to comparatively evaluate the benefits and drawbacks of automated safety features. Nor is the United States Department of Transportation at present in a position to unilaterally take up these charges and make them its own, though organizational transformations of the state apparatus are a possible end result of the public's organization. Following Dewey, the problem is not how to replace a largely-privatized technocratic regime with either a state-run or populist one, but to nurture the types of agency requisite with deliberation about problems whose scales are introduced or remade by transformative technologies. The public is not some metaphysical source of political legitimacy, but a potentially active substance for matching available tools to emergent social problems.

We draw inspiration from Nelkin and Pollak's 1979 portrayal of legitimacy as a substance amenable to institutional intervention, in particular addressing technical challenges through new procedural conditions so that public problems at least remain a subject of active, agonistic deliberation:

> "What then must one do to enhance legitimacy? What kind of procedures would be acceptable to critical groups? To explore the sources of persistent cynicism about procedural reforms, we turn to five questions frequently asked by opposition groups: How are the boundaries of the problem defined? Who participates in the experiment? Who conducts the procedure? What is the distribution of technical expertise? Is there really a choice?

[38] The implications and recommendations are preliminary and will be revisited and tied to their original inspiration from the Hard Choices Framework, as well as the covered accounts of Dewey and Grunig. We hope the WeRobot discussion will help us enrich this part of the paper significantly.

> What can be generalized is not the structure of the experiments, but the conditions that will allow dissenting groups to express their concerns and to communicate effectively with administrative agencies. These conditions include: A 'formula' that gives due weight to social and political factors; appropriate involvement of affected interests; an unbiased management; a fair distribution of expertise; and a real margin of choice. Actually, such procedural conditions are not likely to produce consensus, but they may reduce public mistrust and hostility toward political and administrative institutions in order to allow détente. Our conclusion, in fact, is that détente is a more appropriate and realistic goal.[39]

Building on this vision, we propose a set of basic institutional reforms to address the normative indeterminacies examined in this paper, specifically through interventions on the sources of indeterminacy at stake in AV development. Without endorsing a particular normative vision for the future of public space or transportation, we believe these reforms are necessary for "active détente" with the public, and for preserving the legitimacy of public space as it is remade by automated vehicle fleets. These reforms are described below:

1. **The ODD cannot refer exclusively to technical constraints, and must include constraints on the "safety control structure".** This includes reference to the organization of the road environment, beyond vehicle performance, and the necessary institutional checks and balances to ensure that manufacturers define the ODD with guidance from urban planners as well as the human factors community. ODDs need to be understood as components of public infrastructure that specify a contract mapping observable and controllable features to concrete and verifiable safety guarantees as well as associated liability rules.
2. **Nationwide consumer surveys and outreach marketing must be complemented with concerted engagement of local communities.** This would allow surveyors to index information about AV capabilities within the context that road users are already familiar with. It would also provide channels for members of the public to provide feedback about the "unknown unknowns" of AV deployment, including how a given behavior specification will interact with the incentives of different subpopulations as well as concrete fears or anxieties about performance. Such an initiative, understood in the form of a public interest campaign, would make it possible for affected communities to affirm AV deployment conditions, rather than merely acquiesce to them.
3. **In indeterminate traffic situations, definitions of robustness for safety-critical features must include validation from third parties.** This would help establish new professional norms for deep learning practitioners that are coincident with those of more established fields like civil engineering, which also interface and work closely with the public sector. These norms include accountability for chosen modeling choices, as well as standardized expert qualifications suitable for testimony in legal cases related to negligence. While this may eventually run into issues of

[39] Nelkin D, Pollak M (1979) Public participation in technological decisions: reality or grand illusion? Technol Rev (August/September) 1979:55–64.

intellectual property, what will be needed are forms of standardization and legal rules to align these norms with established juridical procedures, as is now underway with the European Commission's Artificial Intelligence Act[40]). In the long term, external validation and oversight could be integrated into federal or state guidelines for AV deployment certification, aligning modeling procedures with established forms of democratic governance.

Conclusion

While modest, each of these reforms would support an essential component of sociotechnical specification. The common thread across all three is the need for expert professions to take a more active role in organizing and representing distinctive publics as they come into being through the development of automated transportation capabilities. For reasons clearly articulated by Dewey and Grunig, this will require professionals to engage in a complex dance of informing, persuading, and being receptive to the public at distinct stages of its maturation. While difficult, this approach seems clearly better than the present paradigm, in which different strategies of public engagement are utilized either haphazardly or on behalf of organized corporate interests to atomize, isolate, and otherwise disempower the public before it has had a chance to self-determine. So considered, experts' active attention to and engagement with these questions is not just morally desirable, but also an analytic requirement to ensure that control of public space remains in the hands of the agent that it is meant to serve: the public itself.

[40] European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) | Shaping Europe's digital future*. European Commission. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence