When Engineering Meets Economics: AI-Powered Safe and Accountable Autonomous Driving

Xuan Di[a,b,*], Herbert Dawid[c], Gerd Muehlheusser[d]

[a] *Columbia University, Department of Civil Engineering and Engineering Mechanics*
[b] *Data Science Institute, Columbia University*
[c] *Bielefeld University, Department of Business Administration and Economics, and Center for Mathematical Economics*
[d] *University of Hamburg, Department of Economics, and IZA, and CESifo*

Abstract

By now, AI has already affected many aspects of daily life, and its importance can be expected to grow even larger. Relying heavily on AI, autonomous vehicles (AV) are complex engineered systems that can make life and death decisions. Because of their profound impact on society, AVs must be designed and developed to be *safe*, *accountable*, and ultimately *trustworthy* to stakeholders and *responsible* to the society. This paper aims to discuss a pathway to design and develop safe and accountable AVs from the interdisciplinary perspective of engineering, economics and the economic analysis of law. We will primarily discuss different approaches in programming safety rules into AV decision making, and how these approaches fall within the behavior or outcome oriented paradigm. Building on these paradigms, the widely used reinforcement learning approach to train AVs belong to outcome based, while the imitation learning based approach belong to behavior based. Understanding what driving tasks belong to which paradigm would facilitate the design of safety principles to build trust. Also, the distinction between outcome based and behavior based approaches helps to connect the task of training AVs to well-established results from agency theory on how to optimally induce desired actions from human agents. Agency theory also allows us to provide guidance for aligning the interests of different parties in the AV value chain. We also investigate how tort liability can foster the accountability of AVs.

Keywords: Autonomous vehicles (AVs), road safety, accountability of AI systems

1. Introduction

Artificial intelligence (AI) has made significant breakthrough in various fields, including Atari (Mnih et al., 2015), Go game (Silver et al., 2016), Poker (Brown and Sandholm, 2018, 2019), Dota 2 (OpenAI, 2018), StarCraft II (Vinyals et al., 2019), and DALL-E (OpenAI, 2022b). With the booming of chatGPT (OpenAI, 2022a, 2025) and recent DeepSeek (DeepSeek, 2025), AI has seen unprecedentedly widespread adoption. It has been and will perforate every aspect of our daily life, including but not limited to, healthcare decisions, medical diagnosis, product recommendations, university admission and labor recruitment, and more, autonomous driving and legality. In all these applications, AI has demonstrated its great potential in enhancing prediction accuracy, work efficiency, cost effectiveness, and optimal decision making.

Relying on AI in every stage, autonomous vehicles (AV) are complex engineered systems consisting of sensing and perception, object detection and recognition, and planning and decision making. Driving algorithms programmed in AVs could potentially reshape future traffic patterns on road.

As "high-risk AI systems"(Kop, 2021), AVs can make life and death decisions. Unlike other autonomous systems that are tested entirely in laboratories before being deployed in the field, tests of AVs have already

[*] Corresponding author. Tel.: +1 212 853 0435;
Email address: `sharon.di@columbia.edu` (Xuan Di)

come side by side with its development to some extent (WaymoOne). Prior to the maturity of technology and regulation, AVs are already sharing public roads with human drivers[1]. Road users who potentially encounter AVs are part of the natural experiments in the course of perfecting the technology while potentially risking their own lives. As a result, AVs might cause severe, even fatal accidents, especially when they have to interact with humans. Because of its profound impact to the society, how AVs are designed, developed, deployed, and managed is crucial. AVs must be designed and developed to be ultimately *responsible* to society and *trustworthy* to stakeholders.

According to Wing (2021), trustworthiness requires AI to be safe, accountable, fair, and ethical (SAFE). In this paper, we will primarily focus on the former two. While fairness and ethics are important social values in algorithmic programming, safety depicts the technical characteristics that AVs have to fulfill, in design, operation, and management. Furthermore, safety, fairness, and ethics are lower level attributes, while accountability has to be built on the other three. In other words, even with safety, fairness, and ethics in place, if these aspects are violated and do harm, who will be accountable is indispensable to answer. Regulatory bodies leverage the accountability framework to ensure that AI developers and operators take adequate safety measures prior to deployment.

Note that there are a decent amount of technical papers on how to achieve trustworthy AI (Kaur et al., 2022; Kuru, 2022; Li et al., 2023; Kuznietsov et al., 2024). In this paper, we instead focus more on high-level ideas at the intersection of engineering, economics and law, with forward-looking perspectives. Particularly, this paper aims to draw a roadmap and project promising directions for the development of accountable AVs. This review paper is timely in the transition period to an era of full driving automation. We hope it will not only stimulate active conversations and inspire the engineering, economics, and legal communities to rethink the conventional approaches that are developed solely from the perspective of a single discipline, but join forces towards creating accountable AVs and other high-risk AI systems.

1. From an engineering point of view, AVs are complex socio-technical systems equipped with sensors, communication, hardware, and computing powers, driven by AI algorithms. Before implementation, these systems need to pass safety tests, be certified according to certain safety standards, and be held accountable when failure nevertheless occurs.

2. From an economic and legal point of view, AVs are consumer products associated with substantial uncertainty with respect to performance and safety in (mixed) traffic environments. Inducing suitable and safe actions from AVs resembles the task of inducing such behavior from a human driver, a problem studied extensively in the context of *agency theory.* Insights from this theory also provide guidance for designing contracts along the AV value chain which, together with a suitable liability regime and regulatory rules, provide incentives for steering performance and safety of AVs in (socially) desirable way. In law, the area of tort law addresses the issue of legal responsibility (in particular, liability) for harm caused by one party (e.g. an AV) to third parties in the course of an accident. Moreover, AVs could potentially be empowered by human-like AI that may go out of control of human creators and produce undesired outcomes. Thus, we need to identify whom to be held accountable even when safety is assured, or when law is violated to ensure safety guarantees.

1.1. Paper scope

We first focus on the safety principle, which is a proactive process to enforce compliance. Then we move on to accountability, a reactive process to determine who will be hold accountable for after some safety principle is violated or even an accident occurs. In other words, safety rules and principles have to be programmed algorithmically, while accountability guides such programming through providing incentives to behave in a desired way. The research questions we would like to address include:

[1] According to the California Department of Motor Vehicles, by the end of January 2025 AV permit holders had logged over 4 million test miles in California: `https://www.dmv.ca.gov/portal/news-and-media/over-4-million-test-miles-logged-by-autonomous-vehicle-permit-holders-in-california`

1. How do engineers and software developers encode safety objectives in learning algorithms to ensure proactive safety by design?
2. On the top of proactive design, how can insights from economic theory and the economic analysis of the law help improving AV learning algorithms and how can they be used to design contracts along the AV value chain and a legal environment, that facilitates safety compliance?

This paper aims to discuss a pathway to design and develop safe and accountable AVs, thereby taking an interdisciplinary perspectives including the fields of engineering and economics. We will tackle some issues regarding accountability that allegedly hold back the development of AVs. We would also like to understand how each actor, old or new, would face accountability issues at different stages of AV deployment and AI evolution. Future prospects will be provided hoping to inspire researchers, policymakers, lawmakers, regulators, and educators to work hand in hand to resolve unsolved challenges of AI-powered AVs. We will primarily discuss different approaches in programming safety rules into AV decision making, and how these approaches fall within the dichotomy of behavior versus outcome oriented paradigms.

1.2. Paper organization

The organization of the paper is as follows. In Sec. 2, we will introduce AI systems and the two paradigms, *behavior versus outcome based control*, that will facilitate our discussion of safety design principles. In Sec. 3, we will discuss how two types of learning methods, *reinforcement learning and imitation learning*, fall under these two paradigms. Such a dichotomy would provide insights into how each learning method could be adapted to incorporate the safety principle. We then demonstrate how AI algorithms should be programmed to drive an AV to follow other cars safely. Based on findings from agency theory, in Sec. 4, we discuss how behavior and outcome based approaches for inducing (safe) actions from human drivers could help to enforce safety compliance of AVs. In Sec. 5, we discuss the role of liability towards the aim of enhancing AV accountability. Finally, in Sec. 6 we conclude and provide some further prospects to achieve AV safety and accountability.

2. AI-powered trustworthy autonomous driving

Definition 2.1. Artificial intelligence systems (Russell and Norvig, 2023): Artificial Intelligence (AI) is the development of computer systems that are able to perceive their environment and to deliberate as to how to best act in order to achieve their own goals, assuming that the environment contains other agents similar to itself. An AI system is autonomous, adaptive, and interactive, characterized by their autonomy to decide on how to act; ability to adapt by learning from the changes affected in the environments; and how they interact with other agents to coordinate their activities in that environment (Floridi and Sanders, 2004).

As the level of driving automation increases from 0 (no autonomy) to 5 (full autonomy) (SAE, 2018), the levels of autonomy, adaptation, and interaction increase accordingly. In other words, vehicles with a high degree of autonomy have to make decisions more autonomously, adapt to dynamic traffic environments with higher flexibility, and most importantly, be more capable of learning to interact with other road users. Below we will exemplify how various AI algorithms would help an AV learn and adapt to dynamic traffic environments.

An autonomous driving system, like the brain of a car detailed in Fig. 1, consists of (i) perception, (ii) planning, and (iii) control modules, where each module is programmed with certain AI algorithms. For example, in the perception stage, computer vision using (deep) neural networks is commonly used for object detection, segmentation, 3D tracking, pedestrian detection, and simultaneous localization and mapping (Ras et al., 2018). In the planning and control stages, deep reinforcement learning is widely used (Aradi, 2020; Kuutti et al., 2020). In particular, in path planning (Aradi, 2020), AVs or mobile robots are required to navigate a spatial map from its current position to a target location using shortest paths without collision to other static or moving objects. In longitudinal and lateral control (Kuutti et al., 2020), AVs need to select

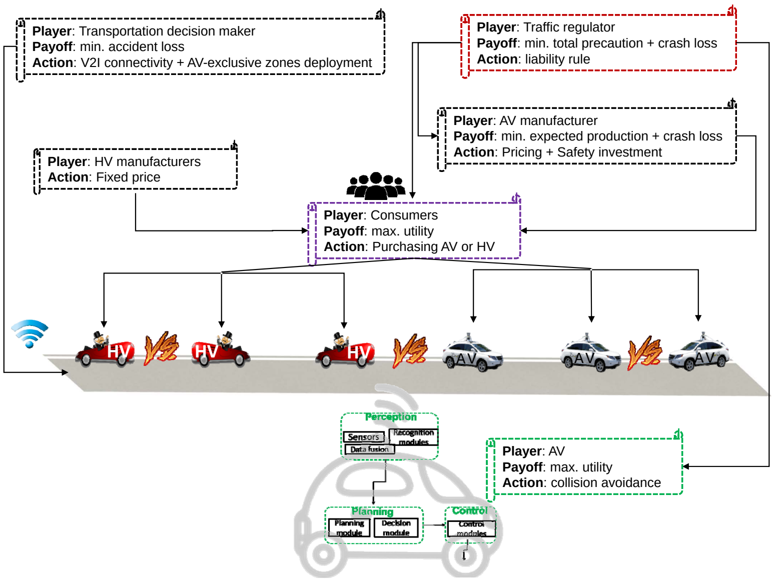

Player: Transportation decision maker
Payoff: min. accident loss
Action: V2I connectivity + AV-exclusive zones deployment
Player: Traffic regulator
Payoff: min. total precaution + crash loss
Action: liability rule
Player: AV manufacturer
Payoff: min. expected production + crash loss
Action: Pricing + Safety investment
Player: HV manufacturers
Action: Fixed price
Player: Consumers
Payoff: max. utility
Action: Purchasing AV or HV
HV
VS
HV
HV
VS
AV
AV
VS
AV
Perception
Sensors
Recognition modules
Data fusion
Planning
Planning module
Decision module
Control
Control modules
Player: AV
Payoff: max. utility
Action: collision avoidance

Figure 1: Transportation Ecosystem. (Each box indicates a player in the traffic system, and lines indicate how actors interact with one another. In the autonomous driving pipeline within an AV control system, the AI algorithms are debriefed in Sec. 2.)

driving acceleration, and driving lanes, while accounting for gaps from vehicles in front or in adjacent lanes, speed differences with these vehicles, and their driving intents. For instance, if one vehicle in the left lane is planning to move to the lane where the ego vehicle is, the ego vehicle may need to re-evaluate its decision to change to the left lane for the moment. Tab. 1 decomposes some driving tasks along the autonomous driving pipeline.

Driving Task	**Sensing and perception**	**Planning and decision making**
Stopping at a Red Light	Use camera to detect the traffic light status	If the light is red, stop the car.
Executing Pre-defined Parking Maneuvers	Use camera and Radar to detect the location of the parking spot and the distance from other cars	Follow a pre-programmed trajectory for parking.
Braking for Pedestrians in a Crosswalk	Use camera or Lidar to detect the presence of a pedestrian	Stop when a pedestrian is detected in a crosswalk.
Maintaining a Safe Following Distance	Use distance sensors like Radar or Lidar to detect cars in front	Maintain a minimum gap from the vehicle ahead.

Table 1: Driving task decomposition

In a nutshell, the major risk involved with AVs, different from other autonomous systems, is the highly dynamic, uncertain, complex mixed traffic environment in which it navigates. Such an environment, with actors such as human drivers, pedestrians, other road users, who could behave erratically or unpredictably, cannot be ignored in the design of an accountable ecosystem, especially when humans would exhibit adaptive behavior day by day, and even develop “moral hazard,” a phenomenon in which one party is incentivized to behave recklessly without bearing the full risk or cost (Pedersen, 2003). When AVs have to adapt to a dynamically changing traffic environment with other adaptive agents, it is challenging to verify and predict the system behavior, potentially leading to unexpected or undesired outcomes. As a result, design principles of safety and accountability have to account for these unique characteristics in autonomous driving systems.

2.1. Safety versus accountability

A trustworthy AV must integrate real-time safety protection mechanism with post-event accountability frameworks to build trust, despite that safety and accountability are interdependent yet distinct characteristics. Safety ensures that an AV functions safely, while accountability requires to justify, trace, and regulate its actions. We first provide formal definitions for these terms.

Definition 2.2.

1. **Safety** (Wing, 2021): the ability to minimize harm by adhering to traffic laws, avoiding collisions, and ensuring reliability in various driving conditions.

2. **Accountability** (Dubber et al., 2020): the requirement for the system to be able to explain and justify its decisions to users and other relevant actors. To ensure accountability, decisions should be derivable from, and explained by, the decision-making mechanisms used. It also requires that the moral values and societal norms that inform the purpose of the system, as well as their operational interpretations, have been elicited in an open way involving all stakeholders.

Safety is a prerequisite for accountability. Safety rules need to be coded into driving algorithms, including collision avoidance, law-abiding, and regulation compliance. On the other hand, accountability is a mechanism to enforce that safety measures are complied with and continuously improved. In particular, when design or oversight is insufficient, legal rules should define how responsibility is apportioned among manufacturers, operators, designers, engineers, insurers, users and third parties.

Ideally, safety and accountability should align. In other words, when an AV is programmed to be safe, it should be accountable, in other words, being able to justify its decisions to stay safe. Sometimes, however, they might conflict, indicating that an AV being safe does not automatically make it accountable. For instance, if an AV avoids a crash but breaks a traffic law in the process, accountability mechanisms must explain and justify its actions. Safety requires flexibility, while accountability demands strict rules and traceability. This inevitably leads to trade-offs. Balancing safety and accountability requires a combination of behavior based (i.e., rule-following) and outcome based (i.e., accident-prevention) control.

2.2. Two paradigms: Behavior versus outcome based approaches

Driving is a complex decision making process with rather involved tasks, which requires taking multiple actions simultaneously. In particular, on the operational level, driving tasks include steering, braking, and acceleration; on the tactical level, they are lane-changing, lane-keeping, car-following, merging and diverging. The strategic level decisions include selecting destinations, departure time, and routes. Nevertheless, traffic is a dynamic environment with constantly moving and interacting road users and objects. A single method will thus not work to ensure safety and build trust. We propose two paradigms to achieve safety and accountability, namely, *behavior based versus outcome based* paradigms. In other words, when it comes to program safety principles and design accountability frameworks, we can decompose each task by its stages and apply different learning methods, based on whether they belong to behavior or outcome based paradigms.

To understand what driving tasks constitute behavior or outcome oriented paradigm, we first need to elaborate on two notions, namely, task *programmability* and outcome *measurability* (Eisenhardt, 1989).

Definition 2.3.

1. **Task programmability** refers to the degree to which appropriate behavior by the agent can be specified in advance.

2. **Outcome measurability** indicates the degree to which crystallized goals can be defined and measured in an allowable amount of time.

Building on these two notions, we can go ahead and map driving tasks onto the paradigms of behavior versus outcome based approaches introduced above. Basic maneuvers, including parallel or backup parking, stop at a stop sign, are easier to measure and program, thus should employ behavioral based control. On the other hand, complex maneuvers such as merging and diverging are hard to program but easier to measure, thus warranting outcome based control.

Behavioral based tasks are easier to program but costly to measure, while outcome based tasks are harder to program but cheaper to measure. The control of precise driving behaviors requires close monitoring by companies, resulting in high cost of measurement, while performance based control simply requires measuring outcomes without measuring detailed behaviors. Accordingly, behavioral based control is rigid and might be unable to adapt to changing environments, while performance based control holds the potential to be adaptive and interactive, but harder to program. The comparison between these two paradigms is summarized in Tab. 2. We would like to remark here that, what driving tasks belong to what paradigm require us to inspect the *programmability* and *measurability* of each task, respectively. However, some driving tasks or maneuvers are complex and might not be easily categorized as one or the other paradigm, and might need both control frameworks. We exemplify tasks in the context of two paradigms in Fig. 2.

Feature	**Behavior based control**	**Outcome based control**
Goal	Following predefined rules	Achieving safe & efficient driving outcomes
Programmability	Explicit programming	Learning from past driving experiences
Measurability	Frequent and accurate sensor inputs	Sparse measurements with outcome performance
Strengths	Predictable, rule-compliant	Adaptive, optimizes for safety & efficiency
Weaknesses	Struggles in novel situations	Harder to interpret decisions

Table 2: Behavior versus outcome based AV control

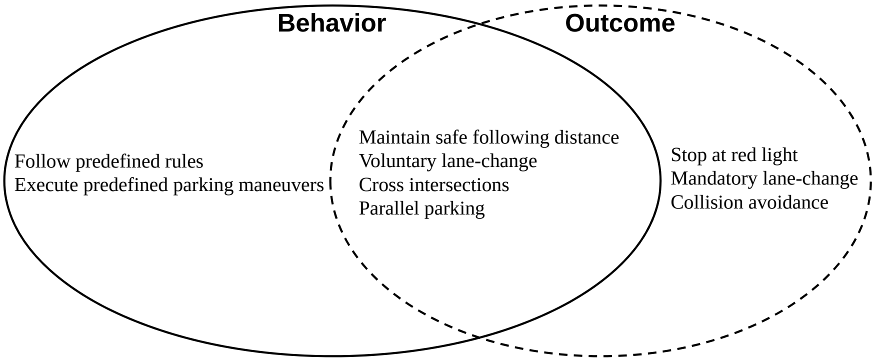


Figure 2: Venn diagram for driving tasks in behavior based and outcome based framework

With these two paradigms defined and understood, we will investigate the safety design philosophy in the context of learning methods and agency theory. Specifically, in Sec. 3, we will discuss how the emerging learning methods can be mapped onto these two paradigms, which would help us better understand how safety should be encoded into AVs' algorithmic decision making processes.

Incentivizing AVs using engineered AI algorithms has become the focus of engineers who design software and hardware systems. When it comes to incentivizing AVs, AI tools are applied at every stage of the system, from planning to decision-making. In the planning stage about trajectory planning, machine learning models are widely used, including end-to-end supervised learning (i.e., from sensed information to decision making), reinforcement learning, and imitation learning. We will discuss how to incorporate safety design principles in these learning schemes.

Following a similar line of thought and inspired by agency theory (see e.g. Laffont and Martimort, 2002), we will then move onto a framework in which someone has to be held accountable if safety principles are violated. To this end, additional actors are introduced, so-called "principals," who are in charge of designing incentive or penalty mechanisms to oversee safety. Accordingly, we will discuss how behavior versus outcome based contracts should be designed to align incentives between principals and agents.

3. Safety by design

AVs are primarily trained using learning methods, in particular reinforcement learning (RL) and imitation learning (IL). In essence, RL methods are outcome oriented. In other words, some reward function is pre-specified and programmers hope that computers could learn a sequence of decisions to optimize such a reward, regardless of whether the learned behaviors are unexpected, as long as the specified reward is optimized. For example, a lot of attention have been devoted to outcome based test (Yu et al., 2024), including collision avoidance and law abiding. Until recently, there are a growing number of studies on behavioral based test (Feng et al., 2020, 2023; Zhang et al., 2024). The IL methods, on the other hand, are behavior-oriented, meaning that rather than stipulating a reward upfront, behavioral trajectories from human experts are provided for an AV to imitate. The goal of AV training is to mimic how humans drive in each state at each time instant. However, behavior based programming and test still face many challenges, in particular, how to assess whether the learned behaviors are accountable and responsible. Below, we will discuss the philosophy of these learning methods and how safety design principles should be customized for each method.

3.1. Reinforcement learning

Reinforcement learning (RL) is a machine learning framework that enables an agent to learn to take a sequence of optimal actions that maximize a cumulative reward (or minimize a cumulative penalty) by interacting with its environment. To "incentivize" AVs to achieve certain goals like collision avoidance, driving comfort, minimization of travel time, a reward function needs to be defined, usually a function of multiple factors combining safety (e.g., no collision), driving efficiency (e.g., to reach a destination as fast as possible), and rule-obeying (e.g., without exceeding speed limit). When this reward is assigned to the host vehicle, namely, it carries a positive value if the host vehicle tries to brake if it is too close, and negative if it does not brake.

A reward could be provided sparsely or densely. A sparse reward is offered at the end when a task is executed, while a dense reward is given to the agent more frequently at every step when an immediate decision needs to be made. The sparse reward is easier to provide, for instance, a lane-change goal could result in a successful (with a reward of 1) or failing (with a reward of 0) action. Sparse rewards, however, could lead to slower learning rate. *Reward shaping* (Ng et al., 1999) is a technique to provide intermediate rewards and guide the learned behaviors of an agent towards desired ones.

In contrast, dense rewards are hard to specify at every time instant. For instance, an AV needs to navigate an urban intersection. A direct goal for the AV is to cross the intersection without colliding to other road users. However, this sparse reward may require a lot more trials for an AV to finally learn how to cross the intersection. Instead, we can assign a reward at every control instant, including headway and speed difference from front cars, lateral distances from cars in adjacent lanes, kinetic energy, passenger comfort, which likely would help better guide the AV to control its lateral and longitudinal accelerations. However, a dense reward could lead to unwanted or inferior behaviors in the learning process, if not properly designed. Thus, how to design a reasonable dense reward remains an engineering challenge.

In the RL framework, there are two options at each step when an AI agent is trained, "exploitation" (which exploits a (local) optimum while interacting with the existing environmental state) and "exploration" (which tries a different solution that may not optimize an immediate incentive but help with a long-term objective). The “exploration” leaves AI agents certain freedom to explore other candidates that may have not been considered in a non-AI approach. If the exploration rate is high, meaning that AI agents explore the environment more than exploit it, it may lead to unexpected solutions that may not be explainable but effective in terms of cumulative rewards. The effect of such exploration leads to so-called “*reward hacking*” (Skalse et al., 2022), where the agent unexpectedly finds a policy associated with a higher reward by exploiting a predefined objective function without genuinely learning the desired behaviors, which deviates from what designers intended. A well-known example in robotics is that a robot designed for ball paddling attempts to lift the racket while keeping the ball resting on it (Kober et al., 2013). Thus, designating a reward function could unintentionally create a shortcut or loophole that can be exploited.

Researchers in AV training discovered a similar outcome that, in a multi-lane traffic, in order to train an AV to control traffic speed, the AV instead learns to drive “zig-zag” across different lanes, which could violate law or cause confusion (Kreidieh et al., 2022). Furthermore, if exploration is out of the developers’ control, the trained AI agents may result in unintended consequences, or “ill will,” if these agents evolve to outsmart their developers.

If we are to create a performance related incentive system in the context of RL, the goal is to encode liability cost, the cost of undesirable behavior such as zig-zagging or the loss borne by the AV on the upper level, so that even when the AV wants to take some actions, for example, to reach a destination sooner by overtaking the cars in front on a single-lane highway, the costs of such an action are so high that it will not be chosen. However, since the reward function is normally a combination of different factors, it can be seen as a “soft” constraint, meaning that the AV can still violate safety if the weight of causing unsafe outcome is smaller than other considerations.

In the RL framework, to impose a “hard” safety constraint, some research (Chen et al., 2019; Bautista-Montesano et al., 2022) proposes a hierarchical control framework, where on the lower level (or inner loop control), the AV solves a RL problem and selects an optimal action that optimizes the reward,

3.2. Imitation learning (IL)

Since rewards could be hard to code manually for various traffic scenarios, another way is to train AVs to mimic professional drivers’ behaviors, which is imitation learning (IL). This approach requires to record how a person drives inside a car with in-vehicle camera recording images facing forward. Then the AV would imitate the sequence of actions based on the input images. When images are fed into this AV’s “brain,” it may or may not extract or parse the semantic information from the images, such as the distance from the car ahead. It may directly take in pixels from these images and imitate when the human demonstrator brake. Thus, the learned driving policy in this circumstance is not interpretable, and sometimes, the AV may learn to brake in a situation when humans do not normally brake. This could happen due to the lack of such scenarios from human demonstrator’s dataset. Thus, an alternative is to let the AV to learn an underlying “incentive” function that drives the demonstrator’s behavior. With this learned incentive function, the AV would select actions by optimizing such an incentive. The potential error is when the AV learns an incentive that deviates from the actual one that drives the behaviors of the demonstrator. If this is the case, the AV could behave unexpectedly and deviate from what the demonstrator has actually demonstrated. Furthermore, a hierarchical control approach (see below) can be used to restrict imitated unsafe actions.

There are two variants of IL, behavioral cloning (BC) and inverse reinforcement learning (IRL). The BC method (Muller et al., 2006; Gu et al., 2020) dodges the difficulty of estimating incentives, because humans may be unable to describe what the exact rewards they actually optimize internally, given that human cognitive processes are complex and sometimes intangible. The Waymo team introduced an augmented BC method, named “ChauffeurNet” (Bansal et al., 2018), which use 30 million expert data and synthesize these data with perturbations that lead to corner cases like collisions or driving off the road. This model is tested in a simulator and also deployed on a real Waymo car by replacing the existing planner module. It shows that when a learned AV learns how to drive from those perturbed datasets offline, the AV is able to deal with extreme cases. More importantly, the learner can respond to causal factors, even though these factors are unobservable to the learner. Since this method is offline learning without letting the AV explore the dynamic environment, its performance in a more complex environment surrounded by human drivers remains unknown. If experts are suboptimal or lack demonstration data in certain scenarios, the imitator trained by BC could fail to learn. In contrast, under IRL the approach is to first learn drivers' intrinsic "incentives" and then infer the actions (Abbeel and Ng, 2004; Shou et al., 2020). Hence, IRL helps the imitator to take actions beyond the training data. Incentives could better infer these actions, which can enable the imitator to behave more optimally than experts, when the experts are suboptimal. However, since incentives are complex to formulate or estimate, researchers normally approximate one's incentive using neural networks, or avoid estimating it but learn policies directly.

3.3. Mapping RL and IL onto two paradigms

We will summarize how variants of RL and IL algorithms are mapped onto the two paradigms (see Tab. 3). In particular, BC belongs to the behavioral based control, because it let an AV mimic behavioral demonstrations from some expert, thus behavioral based. In RL, if we design dense rewards or use reward shaping technique, we hope that the learned driving policies not only optimize some reward, but also shape the learned behavior along the way. The shortcoming is that such a training scheme will not provide flexibility for an AV to adapt to dynamically changing traffic environments. To ensure safety, we have to provide demonstrations that only perform safe actions. On the other hand, IRL and regular RL with sparse rewards belong to the outcome based control, because these methods essentially have to either match a cumulative reward (for IRL) or optimize a single reward (for RL).

Paradigm	**Method**	**Safety design principle**
Behavior based control	Tree search, BC (IL)	safe human driver demonstrations only
Outcome based control	RL with sparse reward	safety in cost (e.g., liability cost)
Hybrid control	IRL	safe human driver demonstrations only + safety set
	RL with dense reward and reward shaping	safety in cost (e.g., intrinsic safety cost + exogenous liability cost)

Table 3: Approach categorization within the two paradigms

3.4. How is an AV programmed to follow others?

Let us exemplify how AVs accelerate or decelerate when following a car head, and avoid collision into the front car. There are several inner working mechanisms, including formal methods and learning methods. The formal methods do not involve prescribing any reward nor incentives, so we briefly discuss their mechanisms, and will primarily focus on RL and IL approaches in which an AV is encoded or needs to learn a reward intrinsically.

3.4.1. Formal methods

Formal methods to control AVs include mathematical models and decision trees, where AVs do not decide autonomously but completely rule-based. For instance, AVs can be programmed based on a mathematical model, so-called car-following model, that reduces the acceleration if its speed is faster than that of the car ahead and the headway is small. A mathematical formula would output a negative acceleration rate proportionally from for example, a widely used car-following model "intelligent driving model (IDM)" (Treiber et al., 2000; Mo et al., 2021).

An alternative approach relies on decision trees. This approach starts from a root node, then checks various criteria sequentially, including the speed difference from the car ahead, the headway, and the existing speed of the host AV. For example, if the AV speed is too fast compared to its car ahead, if the headway is too close, then the action is to decelerate. However, these methods so far fall short because autonomous decision-making is not properly taken into account when AVs can adapt to dynamic environments.

For the industrial practice, Mobileye offers a transparent, formal, and mathematical model to define safety of an AV proactively, namely, “Responsibility-Sensitive Safety (RSS)” (Shalev-Shwartz et al., 2017). An RSS safety rule is a pair $R = (C, P)$, where C is a safety condition and P is a proper response achieving safe driving. A rule assures safety whenever C is satisfied, executing P from that moment leads to safe driving. It is applied to 5 driving scenarios, including maintaining safe distance, do not cut in recklessly, given not taken right of way, be cautious in areas with limited visibility, and avoiding crashes. The rule serves as a precondition to assure safety execution of an intervention, thus, C needs to be *instantly checkable*. In the context of car-following, when $C = (\text{headway} < \text{thread}_{\text{headway}}) \bigcup (\text{speed difference} < \text{thread}_{\text{spd dif}})$, then $P = \text{break}$. To evaluate whether an AV has followed or violated these rules instantly is also a challenge, because the RSS framework is set in a static environment, given that AVs could perceive the environment precisely and traffic scenarios occur with no randomness. When the surrounding human drivers behave unexpectedly or violate rules, the best philosophy for the AV is to avoid collision regardless of the behavior of another actor.

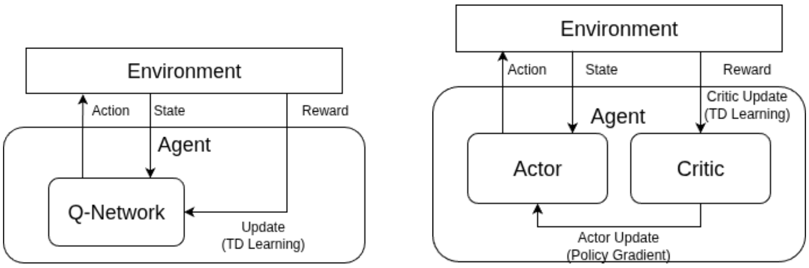
All the aforementioned methods are rule based, as AV do not decide autonomously but rather follow pre-programmed rules. Below we will introduce how learning methods are applied to AVs, which enables the dynamic adaptation of AVs to traffic environments.

3.4.2. Learning methods

To let the AV learn how to follow a car ahead using RL without colliding into this car, a reward needs to be crafted first. The reward could contain terms, including the AV’s own speed, its distance from the leading car, and its relative speed from the leading car (Di and Shi, 2021). If we would like to make the goal of collision avoidance more explicit, we can include another term, a penalty when the distance of two cars is too close under a threshold. Then we can build a computer simulator populated with various traffic environments where the AV tries to explore various actions (i.e., accelerations) following the car in front, with the goal of optimizing a cumulative reward within a prespecified time period. The training process converges, when this AV is capable of selecting an optimal sequence of actions that maximize the prescribed cumulative rewards.

RL algorithms for car-following tasks can be primarily categorized into three types, namely, value based, policy based, and actor-critic (AC) methods that combine both value and policy. Deep Q-Network (DQN) (Masmoudi et al., 2021), a value based method, approximates the Q-function using deep neural networks, has been explored for handling high-dimensional inputs. Proximal Policy Optimization (PPO) (Ni et al., 2024), a policy gradient method, updates the policy parameters in the direction of higher expected rewards. AC algorithms simultaneously learn a policy (i.e., an actor) and a value function (i.e., a critic) to optimize decisions via policy gradients. Deep Deterministic Policy Gradient (DDPG) (Peng et al., 2022; Li and Okhrin, 2023; Hart et al., 2024) utilizes a replay buffer to enhance sample efficiency for the continuous action space. Building upon DDPG, Twin Delayed Deep Deterministic Policy Gradient (TD3) (Bautista-Montesano et al., 2022) introduces an additional critic network and delays the training of actor network to improve stability and performance (see Fig. 3 for illustration on value based and AC based algorithms).

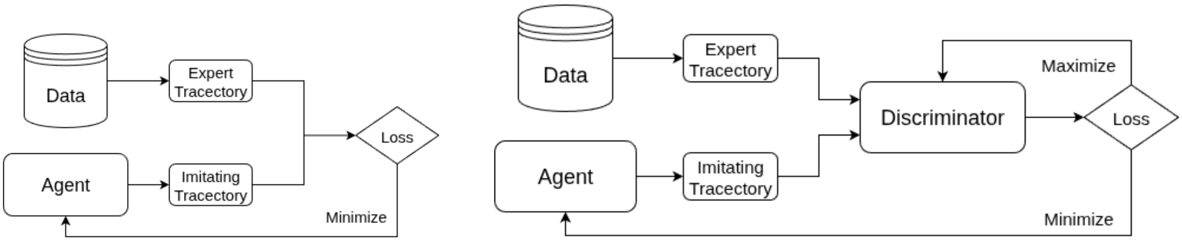
When it comes to the IL method, we need to first collect a set of driving trajectories from human drivers who follow a car in front. The data set could include a set of N human drivers driving histories, denoted as a sequence of state-action pairs, where the state represents the ego car’s position or its relative position to the leader, and the action is the acceleration. We then fit this dataset into a model, with input as the state and output as the action conditional on this state. BC (Dolgov and Michalke, 2020; Li et al., 2024a) is supervised learning that learns a direct mapping from data to action. Instead, IRL (Zhao et al., 2023) first recovers a reward function, and then generates optimal actions based on such a learned reward, with a goal to minimize the distance between the generated actions and the observed ones. Instead of explicitly recovering the reward function, generative adversarial imitation learning (GAIL) (Ruan and Di, 2022a; Ruan et al., 2023a; Tang et al., 2023) trains a discriminator to differentiate expert and agent behaviors, using this signal to train the agent toward producing actions that are indistinguishable from those of the expert (see Fig. 4 for illustration on BC and GAIL algorithms).



(a) Value based RL (i.e, DQN)

(b) AC based RL (i.e., DDPG, TD3)

Figure 3: Schematic framework for RL



(a) Behavior Cloning

(b) Generative Adversarial Imitation Learning

Figure 4: Schematic framework for IL

3.4.3. A hybrid approach

BC or RL with reward shaping could suffer from so-called out-of-distribution test samples, when the test and the training environments differ. On the other hand, RL could potentially lead to reward hacking that even a proper reward might be exploited by the agent to learn some unexpected behavior. Another risk is that a poor reward design or specification could result in unwanted behaviors and outcomes.

To overcome the issues of both paradigms, a hybrid approach has been proposed combing behavior and outcome oriented paradigms. Namely, an AV is first trained using behavior oriented approach from demonstration trajectories, and then outcome based out of rewards, so that the AV can pick up both aspects sequentially. Augmenting the performance of one learning approach using the other learning framework has gained increasing attention (Vecerik et al., 2017; Hester et al., 2018; Song et al., 2023). Policies learned from IL alone are found to often fail to account for safety, despite its power in training human-like behaviors in normal conditions fully captured by demonstration data. Accordingly, a team of Waymo researchers proposes to augment IL with RL of simple rewards (Lu et al., 2023). In particular, the reward consists of a term about the performance of imitation where data is abundant, and another term that directly penalizes safety violation when expert data do not contain such scenarios. This hybrid approach is validated with over 100k miles of urban driving data, and the learned driving policy is more robust on the most challenging scenarios with over 38% reduction in failures.

In addition, a hybrid learning approach could be implemented that focuses on the philosophy of curriculum learning (Anzalone et al., 2022) or meta-learning (Ye et al., 2021). In particular, the training of the AV can start with simple tasks with rule-based models or BC, such as stopping at a stop sign, following speed limit, and other tasks specified by clear regulations and rules. Then the training can gradually move to more complex tasks using IRL or RL with dense rewards, including lane-change, and highway merging and diverge.

3.5. Discussions

How fruitful if one relies on the taxonomy to improve the AV control? The taxonomy of behavior versus outcome based driving tasks would offer conceptual insights into what types of algorithms should be selected for what driving tasks. We summarize how each paradigm guides the safety design principle in the last column of Tab. 3. For the behavior based control framework especially using IL, to ensure safety by design, safe human driver demonstration trajectories need to be provided. For the outcome based control framework relying on RL, a safety cost needs to be incorporated into the reward function, such as liability cost arising from legal constraints. For the hybrid control framework, when IRL is used, both safe demonstrations and safety rewards need to be specified; when RL is employed, safety cost should be explicitly encoded.

This taxonomy provides methodological guidance in terms of what learning algorithms might be desirable for what driving tasks. In particular, the easier it is to define specific behaviors in a driving task, the more suitable the IL is. In contrast, the less costly it is to measure the performance of a driving task, the more suitable it is to use RL. Nevertheless, we would like to clarify that, this taxonomy is coarse, as this provides a philosophical framework, not exact categorization, as indicated in Fig. 2. This is because that, for complex driving tasks that might not be easy to decompose into predefined rules or outcomes, it could not be a good idea to employ one learning method in the training process. The hybrid control offers a pathway to train complex driving tasks in a more coherent framework.

4. Incentivizing AVs: what can we learn from agency theory in economics?

The main goal of AV training is to make the AV behave in a way that is consistent with objectives, such as low accident probability and smooth driving experience. In this sense, there is a strong analogy to the challenge of aligning behavior of human agents, acting on behalf of an organization, with the objectives of that organization. There is a large body of literature in economics dealing with this challenge under the label of *agency theory*. In this section, we briefly discuss this literature and highlight similarities in the basic approach between agency theory and AV learning methods. Furthermore, we raise the question whether insights from agency theory might provide useful ideas for the further development of AV training methods.

4.1. Agency theoretic approaches to deal with moral hazard

In economics, the standard principal-agent models are due to Holmström (1979) and Grossman and Hart (1983). They address the question as to how a principal can optimally steer the behavior of an agent, whose behavior she cannot perfectly observe. The agent acts on the principal's behalf but their objectives are not perfectly aligned (e.g. a manager acting on behalf of the owners of a firm), Problems of this type, where the agent's behavior is not directly observable, are referred to as situations with *moral hazard.* The observable outcomes typically depend in a stochastic way on the agent's action, such that it is neither possible for the principal to infer the exact action from the observable outcome, nor possible for the agent to perfectly predict the effect of its action.

Two main approaches have been proposed for the principal to induce the agent to choose actions well aligned with the principal's objective. First, through the design of a remuneration scheme, which links the payment that the agent receives to the observable outcome. A trade-off arises for the principal because in general the scheme which induces actions of the agent, which are optimally aligned with the principal's interests, is more costly to implement than alternatives which give rise to actions of the agent which are less attractive from the principal's perspective. A large body of literature has analysed in different settings how to design remuneration schemes in order to deal with this trade-off in an optimal way (see e.g., Laffont and Martimort (2002)). A second approach that has been put forward for inducing desired actions of an agent is to try to obtain direct signals of the agent's actions. More precisely, the literature has studied scenarios where the principal can invest in monitoring activities which generate signals of the action chosen by the agent. These signals can be used to adjust the payments to the agent, e.g., an agent can be fined if the signal indicates a deviation of the agent's choice from the foreseen action, and literature in this area has studied how monitoring can be optimally used to steer the agent's behavior (e.g. Mookherjee and Png (1989)). For both approaches results are to a large extent based on the analysis of equilibria in two-stage games, where in the first stage the principal designs the contract and/or the monitoring scheme, and in the second stage the agent chooses an action in light of the decision of the principal in the first stage. A key assumption underlying the vast majority of this work is that the agents always act optimally in the sense that they maximize their utility in light of the remuneration scheme respectively the monitoring activities designed by the principal. The principal is aware of this fact and has perfect or at least imperfect information about the agent's preferences.[2]

The general trade-off between the benefits from agents choosing actions, that are desirable from the principal's perspective, and the costs of implementing such actions, which is the main building block of agency theory, arises in a similar way in the context of AVs. Inducing AVs to behave in a way that is consistent with the goals of AV-producers, AV-users, or more generally society, is costly. A key factor in this respect is the amount of (costly) training data used, which directly induces a trade-off between training costs and AV safety. In spite of this similarity of the general problem addressed by these two fields of research, there are some important differences. Most importantly, the assumption in agency theory that agents have fixed utility functions and can always choose their actions so that this utility function is maximized distinguishes this approach from the challenge of training AVs. The role the reward function plays in the training of AVs by RL algorithms is somehow related to that of the utility function in agency theory, but different from the agent's utility function, the reward function can be determined by the principal (i.e., the designer of the training algorithm) and the agent (i.e., the AV) initially is not able to find the action that maximizes the reward function, but has to be trained to do so. In spite of these differences, the basic challenge of finding effective ways to align behavior of agents with preferences of a principal connects the two approaches.

4.2. Mapping agency theory onto two paradigms

A similarity between agency theory and AV training is that they both follow the distinction between outcome based control and behavior based control. Whereas relying on remuneration schemes for incentivizing agents is outcome based, the use of monitoring devices tries to directly track the behavior of the agent. Hence, there is considerable conceptual similarity between the type of approaches which have been developed to induce the actions of human agents and those developed to train AVs. This raises the question as to whether results in the context of agency theory can provide some insights on how outcome based and behavior based approaches should be combined in an optimal way. However, only few contributions to the agency literature analyze how outcome based and behavior based approaches should be mixed to induce certain actions with minimal costs. For example, Demougin and Fluet (2001) study the optimal mix between the monitoring of behavior and remuneration-based incentives in an agency problem with risk neutral principal and agent, as well as limited liability of the agent. Although the authors derive some interesting general results, such as that the cost for the principal of inducing an action is always larger than the agent's cost of the action, general results about the dependence of the optimal mix on (observable) characteristics of the agent and the agent's task are limited and hard to translate to the context of AV training.

4.3. How is an AV programmed to follow others? Relative performance measures

Focusing on output-based approaches, an important insight from agency theory is that in situations where several agents carry out a similar task, it might be optimal to take into account the relative performance of these agents in the principal's incentive scheme (see e.g. Lazear and Rosen, 1981; Mookherjee, 1984). One standard argument for the use of relative performance measures is that in cases where performances are positively correlated, a good result by one agent indicates a favorable environment. Therefore, the compensation of the other agents should depend negatively on this agent's performance. This reasoning implies that the stronger the positive correlation between outcomes of different agents is, the more useful is the use of relative performance for the principal.[3] A similar argument might be applied to the simultaneous training of a group of AVs, e.g. through reinforcement learning. If they face identical or very similar

[2] Only few recent contributions have considered dynamic principal-agent settings, in which the agent over time can improve his ability to choose actions, that are desirable from the principal's perspective and the principal needs to adjust the remuneration scheme over time in order to facilitate the training of the agent (Pratt, 2015; Garicano and Rayo, 2017; Fudenberg and Rayo, 2019).

[3] As discussed for example in Fleckinger (2012), this simple reasoning does not always hold, and is in particular questionable if the size of the correlation depends on the agents' actions.

scenarios during their training, this should induce a strong positive correlation between outcomes generated by the different AVs.

For example, in the context of car-following tasks, many properties of the environment, such as the aggressiveness of other traffic participants, coincide for different AVs operating in the same region, which generates a positive correlation between outcomes. Hence, relative performance evaluation might be useful in this context. Actually, with respect to human drivers, many companies put stickers on their company cars that encourage traffic participants to call and report observations about the performance of the driver, which can be interpreted as a way to implement relative performance measurement. The insights in agency theory about the merit of the use of relative performance measures in such scenarios, suggest that incorporating the relative performance of an AV, compared to the other AVs in the group, into the reward function might improve learning behavior. Examining in detail the feasibility and prospects of this idea for AV training is beyond the scope of this article. We merely use it as an illustrative example to highlight that, based on the conceptual alignment in the problems addressed by agency theory and AV training approaches, a stronger cross-fertilization of ideas between these two areas could be very productive.

4.4. Agency problems in the AV value chain

Apart from the analogy between steering behavior of human and artificial agents, agency theory is also relevant for the further development of the AV market, since it allows to study and design the interaction between different decision makers along the AV value chain. With a prohibitively high price and intricate regulatory and legislative uncertainties, AVs are most likely to be deployed in the next five years as "robotaxis" (Siddiqui and Bensinger, 2019; Templeton, 2023). Transportation network companies and logistics and trucking industry (Harris, 2022a) have urgently pushed to replace their fleets with shared AVs (WaymoOne; Sumagaysay, 2019; Ohnsman, 2019a,b; Marshall, 2019), due to an estimated cost saving of $1.8 per mile by removing drivers (Siddiqui and Bensinger, 2018). It is anticipated that, even when AVs become more affordable by 2030, owning personal AVs may not be preferable over cheaper shared ride service (Siddiqui and Bensinger, 2018). Robotaxi would be competitive enough to provide cheaper and more convenient on-demand service. Thus, it is reasonable to assume that the business model of AVs is likely to arrive as public fleets instead of private ownership. Accordingly, we believe an earlier business model is to establish a contract between mobility service providers and car manufacturers. Agency theory provides guidance on how to design contracts between mobility operators and AV manufacturers, so that incentives between two parties align.

5. Legal responsibility for accidents involving AVs

5.1. Liability as a key concept of tort law

How to achieve the goal of AV accountability not only includes the tracing, explaining and justifying of actions taken by AVs, but also the issue of legal responsibility for certain types of AV behavior and the consequences thereof (such as causing harm to others in the course of an accident). In this respect, the AV transition might reduce the complexity of the legal design problem in certain ways, for example, by simple removal of human actors from the ecosystem, whose erroneous behavior and incentives otherwise complicate matters, or by making it easier to verify the exact circumstances of an accident ex post in court.

However, unlike deploying robots in a controlled environment, putting AVs on public roads where they interact with other vehicles is a risky activity, especially when these vehicles would make life or death decisions for their passengers or third-party humans around them. It is therefore not surprising that the emergence of AVs has triggered a rich academic debate regarding the issue of *liability* and the eventual need for legal reform in response to decision-making by AVs and autonomous systems in general (see e.g. Marchant and Lindor, 2012; Geistfeld, 2017; Smith, 2017; Eidenmüller and Wagner, 2021; Gless, Silverman and Di, 2022; Buiten, De Streel and Peitz, 2023).

The legal concept of liability is a key element of tort (or accident) law, the area of law capturing situations where (potentially socially desirable) activities by some party might create harm to other, unrelated parties, e.g. in the course of (traffic) accidents (see e.g. Shavell, 2009). Generally speaking, a *liability rule* stipulates how the harm arising in the course of an accident is apportioned among the parties involved. For the context of AVs, this might for example be firms involved in the design, programming, production and distribution of AVs, owners/passengers of AVs, operators of human-driven vehicles (HVs), and third parties such as cyclists or pedestrians.[4]

In the United States and many other countries, tort law has two basic liability regimes, strict liability and fault-based liability (see e.g. Shavell, 2009): Under strict liability, a party is liable for every damage it has caused. By contrast, a fault-based (or negligence) rule has the additional requirement that the party causing the damage must also have violated a given negligence standard (*due care*).

5.2. The economic perspective: Liability as an incentive device

Traditionally, one legal aim when designing liability rules is to ensure that victims are compensated for the harm suffered in the course of an accident. A second aim is to provide incentives for all actors to behave in a socially desirable way *before an accident occurs*, in particular in terms of taking (cost-justified) precautionary measures to reduce the accident risk. This second aim hence introduces also an ex ante perspective into the analysis of liability rules, and it is usually the focus in the economic analysis of liability.[5]

In the following, we provide an overview over some recent formal, game-theoretic studies on AV liability from the field of law & economics. Game theory is a widely used tool in economics, transportation science and many other fields that allows to study the strategic interactions between different actors (see e.g. Fudenberg and Tirole, 1991). For the context of the traffic ecosystem, typical relevant actors to be modeled would be manufacturers of AHs and HVs, human drivers of HVs, regulators, consumers, and other road users such as cyclists or pedestrians. Due to the richness and complexity of the various possible interactions, studies typically focus on a subset of players and interactions.[6] Moreover, most theoretical frameworks of AV liability focus on the comparison of the two core liability regimes in tort law, strict liability and fault-based (or negligence-based) liability, as well es variants thereof.

One class of models thereby considers a scenario with *only AVs* on the streets. Such a scenario can be seen as the endpoint of a potentially long transition process towards autonomous mobility, unlikely to unfold in the near future. In a seminal paper Shavell (2020) proposes a “double liability” rule which holds AV owners strictly liable for all accidents involving their AV. Importantly, however, damages are not paid to parties harmed in the accidents, but to the state. While such a rule appears somewhat unrealistic, Shavell (2020) shows that it would allow to align privately and socially optimal behavior with respect to both precaution (i.e. safety per ride) and activity levels (i.e. overall number of miles driven). Guerra et al. (2022b) study the role of manufacturer residual liability in the sense that an AV manufacturer is hold liable whenever all other parties (e.g. operators and victims) fulfilled their respective due care standards.[7] They show that this liability rule would provide efficient incentives under certain conditions. Schweizer (2024) generalizes the analysis of Shavell (2020) and Guerra et al. (2022b), thereby stressing potential benefits of AVs in making vehicle behavior observable ex post in court, and can hence be taken into account when determining liability. Schweizer (2024) shows that, also in this setting, optimal liability follows from his celebrated *compensation principle* (Schweizer, 2016, 2017).

[4] Another strand of literature has addressed the more fundamental issue that the legal system is historically tailored to human actors. While also firms have entered the picture, there is always a human in the loop (e.g. a firm’s CEO) that can be held legally responsible. However, autonomous systems such as AVs lack such legal personality, which poses some thorny conceptual challenges (see e.g. Wagner, 2019; Eidenmüller and Wagner, 2021; Guerra et al., 2022a; Heine, 2025). Moreover, while we focus on tort law, there also exists research on AV accidents from the perspective of criminal liability (see e.g. Douma and Palodichuk, 2012; Gless et al., 2016).

[5] Note that the two aims are not necessarily mutually consistent. For example, fault-based rules often provide incentives for potential tortfeasors to behave efficiently, but would leave victims uncompensated in case an accident nevertheless occurs, see e.g. Shavell (2009).

[6] See e.g. Elvik (2013) for an early survey of game-theoretic work in transportation science (abstracting from the issue of liability).

[7] To avoid confusion, Guerra et al. (2022b) use the term ’manufacturer residual liability’ somewhat differently than the earlier literature, in which it refers to the shift of (remaining) liability costs to one party once the financial assets of another liable party are depleted (see e.g. Hay and Spier, 2005).

A further strand of game-theoretic literature focuses on *mixed traffic*, i.e. the potentially long transition period in which HVs and AVs co-exist on the streets. It is shown that the legal and regulatory framework plays a key role in mediating the myriad local interactions within the mixed traffic ecosystem, thereby also determining the speed with which the transition toward a world of AV dominance occurs. For example, Chatterjee and Davis (2013) and Chatterjee (2016) analyze how varying the loss share with contributory or comparative negligence would distort human's interaction with AVs. Friedman and Talley (2019) employ a multilateral precaution framework to explore how tort law should adapt to the emergence of AVs in mixed traffic. The potentially optimal legal rules include no-fault, strict liability, and a family of negligence-based rules. None of the above studies assumed that the AV is itself a strategic game player. Di et al. (2020) further study how AV manufacturers could strategically select AVs' safety level using a hierarchical game-theoretical model, in which the lawmaker, AV manufacturer, and human drivers are the high-, mid-, and low-level game players. Chen and Di (2023) model the interaction between AVs and HVs in a specific traffic accident scenario, namely, rear-end crashes, leveraging a matrix game approach. They compare the no-fault, contributory, and comparative liability rules for mixed-traffic platooning.

In all models discussed so far, *AV demand* is either not explicitly modeled or exogenously given. For example, Di et al. (2020) assume that the market penetration of AVs increase monotonically, without accounting for consumers' endogenous demand, nor the effect of liability on market growth. However, in a world of mixed traffic, consumers have a choice between HVs and AVs. The development and dissemination of AVs will therefore not only depend on technological feasibility, but also on how much consumers like them. There is robust empirical (survey) evidence documenting considerable consumer heterogeneity towards AVs both in terms of attitudes willingness to adopt. This literature also identifies crucial factors such as liability, vehicle safety and price, and personal attributes (e.g. Kyriakidis et al., 2015; Shabanpour et al., 2018; Cunningham et al., 2019), which are controlled by manufacturers and policymakers. Against this background, some models have taken the demand side for AV explicitly into account. In Dawid and Muehlheusser (2022), the market shares of AVs and HVs arise endogenously from the players' choices in the game. They compare strict and fault-based liability with respect to firms' incentives to invest in product safety, the timing of AV market introduction, and the AV market penetration over time[8]. De Chiara et al. (2021) consider different liability rules in a static framework in which consumers choose between HVs and AVs, thereby not facing any liability risk when choosing the latter. Dawid et al. (2024) consider four different accident types between AV and HVs, and consumers' vehicle choice affects the mixed traffic composition and hence the prevalence of each accident type. AV safety is determined by the producer (at a cost) and a key factor of of AV demand. A policymaker decides on the liability regime and how much to invest to improve V2I connectivity that reduces the likelihood of accidents in AV-AV interactions. Dawid et al. (2024) show how the liability regime affects AV market penetration, the mixed traffic structure, and overall road safety.

5.3. A practical, measurable framework for AV accountability

From the previous literature, it becomes evident that effective AV accountability requires not only clear legal rules, but also technical features that enable responsibility attribution. To translate these theoretical insights into practical and enforceable standards, it is necessary to identify measurable dimensions of accountability. This section proposes a high-level accountability framework that operationalizes such requirements through four key criteria: transparency, traceability, responsiveness, and explainability. Assessing AV systems along these dimensions can help ensure their accountability to legal, societal, and safety expectations (see e.g. European Commission, 2019; Koopman and Wagner, 2016; ENISA, 2021).

The first criterion, *transparency*, fosters AV accountability by ensuring that decision-making processes are accessible and understandable to relevant stakeholders (European Commission, 2019). Measurable indicators include the availability of decision logs, safety audit reports, and data-sharing protocols that facilitate oversight. *Traceability* takes an ex-post perspective and refers to the ability to reconstruct AV decision pathways and system states during and after incidents such as car accidents. This can be achieved by deploying

[8] In Feess and Muehlheusser (2024), the choice between AVs and HVs depends on behavior of AVs in situations of moral dilemma (swerving in unavoidable accidents), but they do not consider a full-fledged market setting.

detailed event data recorders, system logs, and decision trails that document sensor inputs, actions, and internal states throughout operation (Koopman and Wagner, 2016; ENISA, 2021). In contrast, *responsiveness* addresses the ex-ante perspective: it concerns the capacity for real-time monitoring and intervention, enabling manual oversight or overriding of autonomous decisions to prevent harm. Responsiveness can be measured by the presence (or absence) of manual override controls, fail-safe mechanisms, and real-time alert systems (Koopman and Wagner, 2016; ENISA, 2021). Finally, *explainability* refers to the capacity of AV systems to provide justifications that can be comprehended by regulators, users, investigators, and courts. Measures for explainability include the use of interpretable models and clear explanation reports for decisions made during both normal operation and in incidents (see e.g. Doshi-Velez and Kim, 2017; European Commission, 2019).

These considerations can also be related to the example above of an AV following another car. In the event of an accident arising from the AV's failure to execute appropriate braking or acceleration – whether due to sensor malfunction, flawed control algorithms, or ambiguous traffic conditions – the determination of legal responsibility will likely hinge on the ability to attribute causation and assess the AV's compliance with relevant safety standards. In particular, criteria such as transparency (e.g., accessible decision logs), traceability (e.g., reconstructable sequences of system actions), and explainability (e.g., the ability to justify the AV's real-time choices) are likely to become central in court proceedings. These criteria not only facilitate an objective assessment of whether the vehicle's behavior met the expected standard of care, but also help clarify whether fault lies with the manufacturer, software provider, or other parties.

In summary, these measurable criteria provide a structured basis for evaluating the accountability of AVs. Embedding these dimensions into technical design and regulatory assessment can foster greater trust, facilitate effective oversight, and ultimately support the safe and responsible integration of AVs into society.

6. Challenges and Road Ahead

A large scale availability of AVs could impact how we travel, commute, work, and even live. This paper uses concepts from both engineering and economics that might be useful in overcoming existing challenges regarding AV accountability.

Driving involves a spectrum of tasks ranging from simple to complex, with varying degrees of requirements for safety and reliability. Many times computers could outperform humans easily in some tasks, but worse in others. For example, AI failed cognitive tests that a cognitively normal human would accomplish easily (Zhai et al., 2024; Suchman et al., 2023), while it beat the most intelligent chess player (Silver et al., 2017). Accordingly, algorithmic decisions programmed in AVs only outperform humans in certain scenarios and tasks. Similarly, for the existing automated driving assistance systems, AVs outperform humans in parking and driving in controlled environments like airports, but likely under-perform in tasks like braking for children who suddenly run across the street. To ensure safety and accountability, we should analyze AI algorithms by breaking them down to driving tasks, modules and functionalities, as well as operation domains. Under what conditions and when to employ what AI algorithms need to be examined carefully. Behavior and outcome oriented paradigms shed lights into the conditions under which learning approaches are employed.

Building on the behavior based and outcome based frameworks, we discussed how to guide the safety principle by design. In a nutshell, these frameworks could help better understand what methods suit what driving tasks, in order to ensure AV safety. Driving tasks that belong to behavior based framework should be easy to have defined behaviors and programmability, while those in the outcome based framework should be able to be measured in a cost-effective way. Learning methods falling within the behavior based framework require close monitoring of agents' behavior in the process, while those within the outcome based framework call for crystallized goals and precise measures of achieving these goals. Implementation challenges in learning the AV control still persist, for example, what rewards or remuneration to select that could ensure safety constraints, and how to collect sufficient amounts of safety demonstration samples. In agency theory, how to design rewards to align those between the principal and the agent requires practical examination and constant feedback adjustment.

Several challenges also remain in the context of AV liability. First, attributing responsibility in complex real-world traffic scenarios often requires disentangling the roles of multiple actors—manufacturers, software providers, vehicle owners, and even infrastructure operators—within a highly automated, data-driven environment. Secondly, while measurable criteria such as transparency, traceability, and explainability can assist courts and regulators, the practical implementation of these standards is fraught with technical and legal uncertainties. Moreover, the diversity of legal systems and varying standards across jurisdictions complicates efforts to develop harmonized solutions, potentially resulting in gaps or inconsistencies in accountability. Balancing the need for innovation with societal demands for safety and justice will therefore require ongoing dialogue among engineers, legal scholars, policymakers, and end users. In the next step, we spell out some challenges in more detail.

6.1. Causation as means of accountability

Causal accountability helps to identify who should be responsible for system failure or harm. For example, in U.S. tort law causation is based on a two-step procedure (see e.g. Shavell, 2009): In a first step, it is established whether or not the harm was indeed caused by the respective injurer using a *but-for test.* The second step requires to establish that the injurer was also a *proximate cause*, i.e. the connection between the injurer's action and the harm must not be "too remote". However, distinguishing causation in the real-world is normally challenging and sometimes infeasible, due to the presence of multiple factors that contribute to a certain outcome. For example, a traffic accident might result from the *joint* actions of an AV and the harmed pedestrian.[9] Moreover, a variable could generate a confounding effect, if it has a spurious relation to an effect rather than a real causal relation. For instance, an AV that tries to clone human driving behavior by observing the car dashboard could identify a strong correlation between the brake indicator and the brake action, and mistakenly learn to brake only when the brake light is on (De Haan et al., 2019). Accordingly, the first principle for causation is the time sequence. In other words, if the event A causes an outcome B, A has to happen proceeding B. Confounding occurs when multiple correlated variables lead to a certain outcome (Pearl, 2009). Both A and C lead to B, and A leads to C. In this case, C is possibly a spurious cause to B. For example, normally, icy road conditions are latent to the AV who wishes to imitate a human driver about how to drive safely in these conditions. Without "seeing" ice on the road, a demonstrator's skidding trajectories might be perceived by the imitation learning algorithm as some intended action for no reason, and cause confusion to the AV who aims to imitate such a behavior closely.

Causal inference (Spirtes et al., 2000; Pearl, 2009) addresses the challenge of unobserved confounding bias in the observational data by exploiting causal assumptions about the data-generating mechanisms (commonly through causal graphs and potential outcomes). Causal reasoning plays an important role in forensics. It requires the law enforcement to identify the "chain of evidence." To identify causal relations between inputs and decisions, we could enhance the explainability of an AI model as to whether the model provides an explanation for the process that leads the model from its inputs to outputs. Without a satisfactory interpretability, an AI model's recommendation or decision-making can hardly be trusted. To make AI structurally interpretable, the solutions include encoding explainable components in the architecture design phase (Ras et al., 2018), integrating physics knowledge (Shi et al., 2021a,b,c; Mo et al., 2021, 2022a,b; Mo and Di, 2022), or incorporating causality diagrams (Kocaoglu et al., 2017; Ruan and Di, 2022b; Ruan et al., 2023b, 2024) into the architecture of neural networks. There is a growing body of literature on encoding physics based knowledge into the training phase of AI, for its strength in integrating available scientific hypotheses, i.e., domain knowledge, into machine learning models, i.e., pure data-driven. Such a hybrid structure holds the potential for designing complex and interpretable autonomous systems that comply with physical rules. With the surge of Large Language Models (LLM) and Vision Language Models (VLM), there is also a growing number of literature on generating causal reasoning via LLM or VLM to control AVs (Omeiza et al., 2021; Yang et al., 2023; Tian et al.; Nie et al., 2024; Fu et al., 2024a,b; Li et al., 2024b). This

[9] In law & economics, there exists a large literature analyzing the relationship between causation, negligence standards, the incentives for taking measures to reduce the accident risk, and the calculation of damages (see e.g. Shavell, 1980, 1985; Feess, Muehlheusser and Wohlschlegel, 2011; Schweizer, 2020; Lando and Schweizer, 2021).

could make a breakthrough about whether AVs should be admitted as witness in court (Westbrook, 2017; Seng, 2025), if they are capable of providing causal explanation about their actions prior to accidents.

6.2. Data as means to improve safety and causation

A new challenge and also opportunity brought by AVs lies in the availability of multi-modal multi-fidelity data from various in-vehicle and roadside sensors, including but not limited to camera frames, Radar and Lidar point clouds, phone call records, and driver interaction. Building on these sensor data, rich information could be mined about what involved parties were doing pre-collision, what actions were taken, and what could have been done to avoid it. However, such data types, volumes, velocity and variety are not common evidence admitted to the court, or easily understood by judges, juries, and prosecution agencies, not mentioning machine learning tools used to process, analyze, and interpret these data, which are typically blackbox. Event record device has done this job for regular human-driven cars. With powerful and advanced sensors instrumented on AVs, more personal and private data could be recorded. For example, a Tesla could contain its users' historical phone calls, where they have gone, along with driving videos captured by cameras and other complex sensors (Harris, 2022b; Buquerin and Hof, 2022). On one hand, this rich information holds the potential to tease out the causation after a car accident happens. On the one hand, this triggers the concern about data privacy, despite high utility of the data themselves. When car manufactures have access to more than ever rich data on their customers, it could be challenging for drivers (aka. customers) to defend themselves against the testimony of cars (Gless et al., 2022), particularly when doing so would shift liability to car manufacturers. However, as these technologies are not something commonly admitted in the court, it will be crucial to develop interdisplinary curriculum to educate the next-generation of engineers and law students to grasp the technological knowledge and equip themselves with know-hows to tackle challenges induced by AI.

However, privacy and proprietary protection could come in the way. Blockchain is one approach to store these highly private data from multiple cars using a growing list of blocks linked in chronological order (Nakamoto, 2008). When cars are connected via a distributed peer-to-peer network, they adhere to preserving explaining the elements of anonymity, non-tampering, traceability, and transparency, so that it could be justifiable for the police or investigation team to reconstruct a car accident using the data as legal evidence (Tyagi et al., 2022; Zhu et al., 2024).

6.3. Education and public participation

Education and participation is crucial to ensure the governance of AI for AVs. Unlike human actors, whose after-the-fact testimony can be subject to errors and biases of recollection, AVs are also effective data gathering agents, and the collective data harvested by AVs can make evidence production more transparent, if regulation of data sharing is properly design. In a Columbia lecture taught by the first author and assisted by two law experts in Spring 2023, a mock court was implemented for the first Uber fatality case (DeArman, 2019), in which students played roles as judge, jury, plaintiff, defendant, and attorneys. Prior to the lecture, students surveyed technical details of how such an accident would have happened. Such an activity has improved students understanding and capability of analyzing legal cases, which stimulated students' interest in equipping themselves with interdisciplinary knowledge at the intersection of engineering, law, and economics. Policymakers, lawyers, engineers, and data scientists should join the force to overcome obstacles in data collection, analysis, and provision for legal use, and rethink the curriculum designed nowadays to train engineers, data scientists, lawyers, and economists.

To ensure safety, we do not only need proactive design, but also reactive regulatory mechanisms, including record keeping, oversight, certification, enforcement, and compliance. The instruments to achieve these mechanisms include black box recorders, explainable AI, and post-incident analysis to determine responsibility. We decide not to delve into these mechanisms in this paper, as each could be a whole topic in its own right.

Acknowledgments

We gratefully acknowledge the hospitality by the Center of Interdisciplinary Research (ZiF) at Bielefeld University (Research Group "Economic and Legal Challenges in the Advent of Smart Products") and financial support by the Deutsche Forschungsgemeinschaft (Grant SFB 1283/2 2021–317210226). We would also like to thank Professor Sabine Gless from University of Basel for her insightful discussions on liability, causality, and data privacy, as well as her assistance in Di's class for the mock court trial. We would also like to thank Professor Eric L. Talley from Columbia University for his inspiring discussions. Di would like to thank the support of National Science Foundation CAREER under award number CMMI-1943998.

References

Abbeel, P., Ng, A.Y., 2004. Apprenticeship learning via inverse reinforcement learning, in: Proceedings of the twenty-first international conference on Machine learning (ICML), p. 1.

Anzalone, L., Barra, P., Barra, S., Castiglione, A., Nappi, M., 2022. An end-to-end curriculum learning approach for autonomous driving scenarios. IEEE Transactions on Intelligent Transportation Systems 23, 19817–19826.

Aradi, S., 2020. Survey of deep reinforcement learning for motion planning of autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems 23, 740–759.

Bansal, M., Krizhevsky, A., Ogale, A., 2018. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. arXiv preprint arXiv:1812.03079 .

Bautista-Montesano, R., Galluzzi, R., Ruan, K., Fu, Y., Di, X., 2022. Autonomous navigation at unsignalized intersections: A coupled reinforcement learning and model predictive control approach. Transportation research part C: emerging technologies 139, 103662.

Brown, N., Sandholm, T., 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. Science 359, 418–424.

Brown, N., Sandholm, T., 2019. Superhuman AI for multiplayer poker. Science 365, 885–890.

Buiten, M., De Streel, A., Peitz, M., 2023. The law and economics of ai liability. Computer Law & Security Review 48, 105794.

Buquerin, K.G., Hof, H.J., 2022. Digital forensics investigation of the tesla autopilot file system,", in: SECURWARE 2022, The Sixteenth International Conference on Emerging Security Information, Systems and Technologies, pp. 82–87.

Chatterjee, I., 2016. Understanding Driver Contributions to Rear-End Crashes on Congested Freeways and their Implications for Future Safety Measures. Ph.D. thesis. University of Minnesota.

Chatterjee, I., Davis, G., 2013. Evolutionary game theoretic approach to rear-end events on congested freeway. Transportation Research Record: Journal of the Transportation Research Board , 121–127.

Chen, J., Yuan, B., Tomizuka, M., 2019. Deep imitation learning for autonomous driving in generic urban scenarios with enhanced safety, in: arXiv preprint arXiv:1903.00640.

Chen, X., Di, X., 2023. Legal framework for rear-end crashes in mixed-traffic platooning: A matrix game approach. Future Transportation 3, 417–428.

Cunningham, M., Regan, M., Horberry, T., Weeratunga, K., Dixit, V., 2019. Public opinion about automated vehicles in Australia: Results from a large-scale national survey. Transportation Research Part A: Policy and Practice 129, 1–18.

Dawid, H., Di, X., Kort, P.M., Muehlheusser, G., 2024. Autonomous vehicles policy and safety investment: an equilibrium analysis with endogenous demand. Transportation research part B: methodological 182, 102908.

Dawid, H., Muehlheusser, G., 2022. Smart products: Liability, investments in product safety, and the timing of market introduction. Journal of Economic Dynamics and Control 134, 104288.

De Chiara, A., Elizalde, I., Manna, E., Segura-Moreiras, A., 2021. Car accidents in the age of robots. International Review of Law and Economics 68, 106022.

De Haan, P., Jayaraman, D., Levine, S., 2019. Causal confusion in imitation learning. Advances in neural information processing systems 32.

DeArman, A., 2019. The wild, wild west: A case study of self-driving vehicle testing in arizona. Ariz. L. Rev. 61, 983.

DeepSeek, 2025. Deepseek ai (v3) [large language model]. `https://www.deepseek.com/`.

Demougin, D., Fluet, C., 2001. Monitoring versus incentives. European Economic Review 45, 1741–1764.

Di, X., Chen, X., Talley, E., 2020. Liability design for autonomous vehicles and human-driven vehicles: A hierarchical game-theoretic approach. Transportation Research Part C: Emerging Technologies 118, 102710.

Di, X., Shi, R., 2021. A survey on autonomous vehicle control in the era of mixed-autonomy: From physics-based to AI-guided driving policy learning. Transportation Research Part C: Emerging Technologies 125, 103008.

Dolgov, M., Michalke, T., 2020. Mono-video deep adaptive cruise control in the image space via behavior cloning, in: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 1–6.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 .

Douma, F., Palodichuk, S., 2012. Criminal liability issues created by autonomous vehicles. Santa Clara Law Review 52, 1157.

Dubber, M.D., Pasquale, F., Das, S., 2020. The Oxford handbook of ethics of AI. Oxford Handbooks.

Eidenmüller, H., Wagner, G., 2021. Law by Algorithm. Mohr Siebeck, Tübingen.

Eisenhardt, K.M., 1989. Agency theory: An assessment and review. Academy of management review 14, 57–74.

Elvik, R., 2013. A review of game-theoretic models of road user behaviour. Accident; analysis and prevention 62.

ENISA, 2021. European Union Agency for Cybersecurity: Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Vehicles. Technical Report. ENISA. URL: https://www.enisa.europa.eu/publications/cybersecurity-challenges-in-the-uptake-of-ai-in-avs.

European Commission, 2019. High-level expert group on ai: Ethics guidelines for trustworthy ai. Available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Feess, E., Muehlheusser, G., 2024. Autonomous vehicles: moral dilemmas and adoption incentives. Transportation research part B: methodological 181, 102894.

Feess, E., Muehlheusser, G., Wohlschlegel, A., 2011. Screening in courts: on the joint use of negligence and causation standards. The Journal of Law, Economics, & Organization 27, 350–375.

Feng, S., Feng, Y., Yu, C., Zhang, Y., Liu, H.X., 2020. Testing scenario library generation for connected and automated vehicles, part i: Methodology. IEEE Transactions on Intelligent Transportation Systems 22, 1573–1582.

Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., Liu, H.X., 2023. Dense reinforcement learning for safety validation of autonomous vehicles. Nature 615, 620–627.

Fleckinger, P., 2012. Correlation and relative performance evaluation. Journal of Economic Theory 147, 93–117.

Floridi, L., Sanders, J.W., 2004. On the morality of artificial agents. Minds and machines 14, 349–379.

Friedman, E., Talley, E., 2019. Automatorts: How should accident law adapt to autonomous vehicles? Lessons from Law & Economics. Columbia University, mimeo .

Fu, Y., Jain, A., Chen, X., Mo, Z., Di, X., 2024a. Drivegenvlm: Real-world video generation for vision language model based autonomous driving, in: 2024 IEEE International Automated Vehicle Validation Conference (IAVVC), IEEE. pp. 1–6.

Fu, Y., Li, Y., Di, X., 2024b. Gendds: Generating diverse driving video scenarios with prompt-to-video generative model , 819–824.

Fudenberg, D., Rayo, L., 2019. Training and effort dynamics in apprenticeship. American Economic Review 109, 3780–3812.

Fudenberg, D., Tirole, J., 1991. Game theory. MIT Press.

Garicano, L., Rayo, L., 2017. Relational knowledge transfers. American Economic Review 107, 2695–2730.

Geistfeld, M., 2017. A roadmap for autonomous vehicles: State tort liability, automobile insurance, and federal safety regulation. California Law Review 105, 1611–1694.

Gless, S., Silverman, E., Di, X., 2022. Ca (r) veat emptor: Crowdsourcing data to challenge the testimony of in-car technology. JURIMETRICS 62, 285–302.

Gless, S., Silverman, E., Weigend, T., 2016. If robots cause harm, who is to blame? Self-driving cars and criminal liability. New Criminal Law Review 19, 412–436.

Grossman, S.J., Hart, O., 1983. An analysis of the principal agent problem. Econometrica 51, 7–46.

Gu, Z., Li, Z., Di, X., Shi, R., 2020. An lstm-based autonomous driving model using a waymo open dataset. Applied Sciences 10, 2046.

Guerra, A., Parisi, F., Pi, D., 2022a. Liability for robots i: legal challenges. Journal of Institutional Economics 18, 331–343.

Guerra, A., Parisi, F., Pi, D., 2022b. Liability for robots ii: an economic analysis. Journal of Institutional Economics 18, 553–568.

Harris, M., 2022a. Behind the scenes of Waymo's worst automated truck crash. https://techcrunch.com/2022/07/01/behind-the-scenes-of-waymos-worst-automated-truck-crash/. [Online; accessed 01.09.2023].

Harris, M., 2022b. Who actually owns tesla's data? the company, says the company - but other interpretations persist. IEEE Spectrum .

Hart, F., Okhrin, O., Treiber, M., 2024. Towards robust car-following based on deep reinforcement learning. Transportation research part C: emerging technologies 159, 104486.

Hay, B., Spier, K.E., 2005. Manufacturer liability for harms caused by consumers to others. American Economic Review 95, 1700–1711.

Heine, K., 2025. Autonomous decision-making as a challenge for legal research. Journal of Institutional and Theoretical Economics (JITE) forthcoming.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., et al., 2018. Deep q-learning from demonstrations, in: Proceedings of the AAAI conference on artificial intelligence.

Holmström, B., 1979. Moral hazard and observability. The Bell Journal of Economics 10, 74–91.

Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A., 2022. Trustworthy artificial intelligence: a review. ACM computing surveys (CSUR) 55, 1–38.

Kober, J., Bagnell, J.A., Peters, J., 2013. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research 32, 1238–1274.

Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S., 2017. Causalgan: Learning causal implicit generative models with adversarial training. arXiv preprint arXiv:1709.02023 .

Koopman, P., Wagner, M., 2016. Challenges in autonomous vehicle testing and validation. SAE International Journal of Transportation Safety 4, 15–24. doi:10.4271/2016-01-0128.

Kop, M., 2021. Eu artificial intelligence act: The european approach to ai, Stanford-Vienna Transatlantic Technology Law Forum, Transatlantic Antitrust.

Kreidieh, A.R., Zhao, Y., Parajuli, S., Bayen, A.M., 2022. Learning generalizable multi-lane mixed-autonomy behaviors in single lane representations of traffic, in: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, pp. 1663–1665.

Kuru, K., 2022. Trustfsdv: Framework for building and maintaining trust in self-driving vehicles. IEEE Access 10, 82814–82833.

Kuutti, S., Bowden, R., Jin, Y., Barber, P., Fallah, S., 2020. A survey of deep learning applications to autonomous vehicle control. IEEE Transactions on Intelligent Transportation Systems 22, 712–733.

Kuznietsov, A., Gyevnar, B., Wang, C., Peters, S., Albrecht, S.V., 2024. Explainable ai for safe and trustworthy autonomous driving: a systematic review. IEEE Transactions on Intelligent Transportation Systems .

Kyriakidis, M., Happee, R., de Winter, J., 2015. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. Transportation Research Part F: Traffic Psychology and Behaviour 32, 127–140.

Laffont, J.J., Martimort, D., 2002. The Theory of Incentives: The Principal-Agent Model. Princeton University Press.

Lando, H., Schweizer, U., 2021. Causation and the incentives of multiple injurers. International Review of Law and Economics 68, 106026.

Lazear, E., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. Journal of Political Economy 89, 841–864.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B., 2023. Trustworthy ai: From principles to practices. ACM Computing Surveys 55, 1–46.

Li, D., Okhrin, O., 2023. Modified ddpg car-following model with a real-world human driving experience with carla simulator. Transportation research part C: emerging technologies 147, 103987.

Li, M., Li, Z., Cao, Z., 2024a. Enhancing car-following performance in traffic oscillations using expert demonstration reinforcement learning. IEEE Transactions on Intelligent Transportation Systems .

Li, Y., Mo, Z., Di, X., 2024b. Safeaug: Safety-critical driving data augmentation from naturalistic datasets, in: 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 3251–3256.

Lu, Y., Fu, J., Tucker, G., Pan, X., Bronstein, E., Roelofs, R., Sapp, B., White, B., Faust, A., Whiteson, S., et al., 2023. Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios, in: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 7553–7560.

Marchant, G.E., Lindor, R.A., 2012. The coming collision between autonomous vehicles and the liability system. Santa Clara L. Rev. 52, 1321.

Marshall, A., 2019. 1 year after uber's fatal crash, robocars carry on quietly. https://www.wired.com/story/uber-crash-elaine-herzberg-anniversary-safety-self-driving/. [Online; accessed 07.18.2019].

Masmoudi, M., Friji, H., Ghazzai, H., Massoud, Y., 2021. A reinforcement learning framework for video frame-based autonomous car-following. IEEE Open Journal of Intelligent Transportation Systems 2, 111–127.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.

Mo, Z., Di, X., 2022. Uncertainty quantification of car-following behaviors: Physics-informed generative adversarial networks, in: the 28th ACM SIGKDD in conjunction with the 11th International Workshop on Urban Computing (UrbComp2022).

Mo, Z., Fu, Y., Di, X., 2022a. Quantifying uncertainty in traffic state estimation using generative adversarial networks, in: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 2769–2774.

Mo, Z., Fu, Y., Xu, D., Di, X., 2022b. Trafficflowgan: Physics-informed flow based generative adversarial network for uncertainty quantification. European Conference on Machine Learning and Data Mining (ECML PKDD) .

Mo, Z., Shi, R., Di, X., 2021. A physics-informed deep learning paradigm for car-following models. Transportation research part C: emerging technologies 130, 103240.

Mookherjee, D., 1984. Optimal incentive schemes with many agents. The Review of Economic Studies 51, 433–446.

Mookherjee, D., Png, I., 1989. Optimal auditing, insurance, and redistribution. The Quarterly Journal of Economics 104, 399–415.

Muller, U., Ben, J., Cosatto, E., Flepp, B., Cun, Y.L., 2006. Off-road obstacle avoidance through end-to-end learning, in: Advances in neural information processing systems, pp. 739–746.

Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system .

Ng, A.Y., Harada, D., Russell, S., 1999. Policy invariance under reward transformations: Theory and application to reward shaping, in: Icml, pp. 278–287.

Ni, Y.C., Knoop, V.L., Kooij, J.F., Van Arem, B., 2024. Adaptive cruise control utilizing noisy multi-leader measurements: A learning-based approach. IEEE Open Journal of Intelligent Transportation Systems .

Nie, M., Peng, R., Wang, C., Cai, X., Han, J., Xu, H., Zhang, L., 2024. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving, in: European Conference on Computer Vision, Springer. pp. 292–308.

Ohnsman, A., 2019a. Lyft adding self-driving waymo vehicles for riders in phoenix. https://www.forbes.com/sites/alanohnsman/2019/05/07/lyft-adding-self-driving-waymo-vehicles-for-phoenix-area-riders/#3c37832951be. [Online; accessed 07.18.2019].

Ohnsman, A., 2019b. Self-driving in sin city: Lyft and aptiv notch 50,000 robotaxi rides in las vegas program. https://www.forbes.com/sites/alanohnsman/2019/05/31/self-driving-in-sin-city-aptiv-lyft-notch-50000-robo-taxi-rides-in-las-vegas-pilot/#59d3d1965233. [Online; accessed 07.18.2019].

Omeiza, D., Webb, H., Jirotka, M., Kunze, L., 2021. Explanations in autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems 23, 10142–10162.

OpenAI, 2018. Openai five. https://blog.openai.com/openai-five/.

OpenAI, 2022a. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/.

OpenAI, 2022b. Dall-e: Creating images from text. https://openai.com/blog/dall-e/.

OpenAI, 2025. Chatgpt (v4) [large language model]. https://chat.openai.com/chat.

Pearl, J., 2009. Causality. Cambridge university press.

Pedersen, P.A., 2003. Moral hazard in traffic games. Journal of Transport Economics and Policy (JTEP) 37, 47–68.

Peng, J., Zhang, S., Zhou, Y., Li, Z., 2022. An integrated model for autonomous speed and lane change decision-making based on deep reinforcement learning. IEEE Transactions on Intelligent Transportation Systems 23, 21848–21860.

Pratt, J., 2015. Dynamic contracts and learning by doing. Mathematics and Financial Economics 19, 169–193.

Ras, G., van Gerven, M., Haselager, P., 2018. Explanation methods in deep learning: Users, values, concerns and challenges. Explainable and Interpretable Models in Computer Vision and Machine Learning , 19–36.

Ruan, K., Di, X., 2022a. Learning human driving behaviors with sequential causal imitation learning. the 36th AAAI Conference on Artificial Intelligence .

Ruan, K., Di, X., 2022b. Learning human driving behaviors with sequential causal imitation learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4583–4592.

Ruan, K., Zhang, J., Di, X., Bareinboim, E., 2023a. Causal imitation learning via inverse reinforcement learning, in: The Eleventh International Conference on Learning Representations.

Ruan, K., Zhang, J., Di, X., Bareinboim, E., 2023b. Causal imitation learning via inverse reinforcement learning, in: The Eleventh International Conference on Learning Representations.

Ruan, K., Zhang, J., Di, X., Bareinboim, E., 2024. Causal imitation for markov decision processes: A partial identification approach. Advances in Neural Information Processing Systems 37, 87592–87620.

Russell, S., Norvig, P., 2023. Artificial intelligence a modern approach.

SAE, 2018. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (j3016_201806).

Schweizer, U., 2016. Efficient incentives from obligation law and the compensation principle. International Review of Law and Economics 45, 54–62.

Schweizer, U., 2017. Efficient compensation: Lessons from civil liability. Journal of Institutional and Theoretical Economics (JITE) , 54–70.

Schweizer, U., 2020. But-for causation and the implementability of compensatory damages rules. The Journal of Law, Economics, and Organization 36, 231–254.

Schweizer, U., 2024. Liability for accidents between road users whose activity levels are verifiable. Journal of Institutional and Theoretical Economics (JITE) forthcoming.

Seng, D.K.B., 2025. 'to admit or not to admit': That is the question for ai evidence essays in honour of professor tan yock lin (2025). Available at SSRN 5184567 .

Shabanpour, R., Golshani, N., Shamshiripour, A., Mohammadian, A.K., 2018. Eliciting preferences for adoption of fully automated vehicles using best-worst analysis. Transportation Research Part C: Emerging Technologies 93, 463–478.

Shalev-Shwartz, S., Shammah, S., Shashua, A., 2017. On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374 .

Shavell, S., 1980. An analysis of causation and the scope of liability in the law of torts. The Journal of Legal Studies 9, 463–516.

Shavell, S., 1985. Uncertainty over causation and the determination of civil liability. The Journal of Law and Economics 28, 587–609.

Shavell, S., 2009. Economic analysis of accident law. Harvard University Press.

Shavell, S., 2020. On the redesign of accident liability for the world of autonomous vehicles. The Journal of Legal Studies 49, 243–285.

Shi, R., Mo, Z., Di, X., 2021a. Physics informed deep learning for traffic state estimation: A hybrid paradigm informed by second-order traffic models, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 540–547.

Shi, R., Mo, Z., Huang, K., Di, X., Du, Q., 2021b. Physics-informed deep learning for traffic state estimation. arXiv preprint arXiv:2101.06580 .

Shi, R., Mo, Z., Huang, K., Di, X., Du, Q., 2021c. A physics-informed deep learning paradigm for traffic state and fundamental diagram estimation. IEEE Transactions on Intelligent Transportation Systems .

Shou, Z., Di, X., Ye, J., Zhu, H., Zhang, H., Hampshire, R., 2020. Optimal passenger-seeking policies on e-hailing platforms using markov decision process and imitation learning. Transportation Research Part C: Emerging Technologies 111, 91–113.

Siddiqui, F., Bensinger, G., 2018. Why driverless cars will mostly be shared, not owned. https://www.economist.com/the-economist-explains/2018/03/05/why-driverless-cars-will-mostly-be-shared-not-owned. [Online; accessed 07.18.2019].

Siddiqui, F., Bensinger, G., 2019. As IPO soars, can uber and lyft survive long enough to replace their drivers with computers? https://www.washingtonpost.com/technology/2019/03/29/even-with-ipo-billions-can-uber-lyft-survive-long-enough-replace-their-drivers-with-machines/?utm_term=.671e9b2f20f8. [Online; accessed 07.18.2019].

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of go with deep neural networks and tree search. nature 529, 484.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al., 2017. Mastering the game of go without human knowledge. nature 550, 354–359.

Skalse, J., Howe, N., Krasheninnikov, D., Krueger, D., 2022. Defining and characterizing reward gaming. Advances in Neural Information Processing Systems 35, 9460–9471.

Smith, B.W., 2017. Automated driving and product liability. Mich. St. L. Rev. , 1.

Song, D., Zhu, B., Zhao, J., Han, J., Chen, Z., 2023. Personalized car-following control based on a hybrid of reinforcement learning and supervised learning. IEEE Transactions on Intelligent Transportation Systems 24, 6014–6029.

Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D., 2000. Causation, prediction, and search. MIT press.

Suchman, K., Garg, S., Trindade, A.J., 2023. Chat generative pretrained transformer fails the multiple-choice american college of gastroenterology self-assessment test. Official journal of the American College of Gastroenterology— ACG 118, 2280–2282.

Sumagaysay, L., 2019. California greenlights waymo taxi test program. `https://www.ttnews.com/articles/california-greenlights-waymo-taxi-test-program`. [Online; accessed 07.18.2019].

Tang, X., Yuan, K., Li, S., Yang, S., Zhou, Z., Huang, Y., 2023. Personalized decision-making and control for automated vehicles based on generative adversarial imitation learning, in: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 4806–4812.

Templeton, B., 2023. At-fault robotaxi accidents for waymo, pony.ai, (not) olli and what they mean for the future. `https://www.forbes.com/sites/bradtempleton/2021/12/23/at-fault-robotaxi-accidents-for-waymo-ponyai-olli-and-what-they-mean-for-the-future/?sh=270e25a3432f`. [Online; accessed 1.16.2023].

Tian, H., Reddy, K., Feng, Y., Quddus, M., Demiris, Y., Angeloudis, P., . Large (vision) language models for autonomous vehicles: Current trends and future directions. Authorea Preprints .

Treiber, M., Hennecke, A., Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. Physical review E 62, 1805.

Tyagi, R., Sharma, S., Mohan, S., 2022. Blockchain enabled intelligent digital forensics system for autonomous connected vehicles, in: 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), IEEE. pp. 1–6.

Vecerik, M., Hester, T., Scholz, J., Wang, F., Pietquin, O., Piot, B., Heess, N., Rothörl, T., Lampe, T., Riedmiller, M., 2017. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. arXiv preprint arXiv:1707.08817 .

Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J.P., Jaderberg, M., Vezhnevets, A.S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T.L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., Silver, D., 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature 575, 350–354.

Wagner, G., 2019. Robot liability, in: Liability for artificial intelligence and the internet of things, pp. 25–62.

WaymoOne, . Waymo one program. `https://waymo.com/`. [Online; accessed 07.18.2019].

Westbrook, C.W., 2017. The google made me do it: the complexity of criminal liability in the age of autonomous vehicles. Mich. St. L. Rev. , 97.

Wing, J.M., 2021. Trustworthy ai. Communications of the ACM 64, 64–71.

Yang, Z., Jia, X., Li, H., Yan, J., 2023. Llm4drive: A survey of large language models for autonomous driving. arXiv preprint arXiv:2311.01043 .

Ye, F., Wang, P., Chan, C.Y., Zhang, J., 2021. Meta reinforcement learning-based lane change strategy for autonomous vehicles, in: 2021 IEEE Intelligent Vehicles Symposium (IV), IEEE. pp. 223–230.

Yu, W., Zhao, C., Wang, H., Liu, J., Ma, X., Yang, Y., Li, J., Wang, W., Hu, X., Zhao, D., 2024. Online legal driving behavior monitoring for self-driving vehicles. Nature communications 15, 408.

Zhai, X., Nyaaba, M., Ma, W., 2024. Can generative ai and chatgpt outperform humans on cognitive-demanding problem-solving tasks in science? Science & Education , 1–22.

Zhang, X., Khastgir, S., Tiele, J.K., Takenaka, K., Hayakawa, T., Jennings, P., 2024. Odd and behavior based scenario generation for automated driving systems. IEEE Access 12, 10652–10663.

Zhao, Z., Liao, X., Abdelraouf, A., Han, K., Gupta, R., Barth, M.J., Wu, G., 2023. Inverse reinforcement learning and gaussian process regression-based real-time framework for personalized adaptive cruise control, in: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 4428–4435.

Zhu, C., Hu, J., Wu, J., Long, C., Si, X., 2024. A blockchain-based accident forensics system for smart connected vehicles, in: Proceedings of the 2024 6th Blockchain and Internet of Things Conference, pp. 128–136.