Received: 13 September 2022 | Revised: 9 March 2023 | Accepted: 4 April 2023

DOI: 10.1002/asmb.2765

DISCUSSION PAPER

WILEY

Some statistical challenges in automated driving systems

William N. Caballero[1] | **David Rios Insua**[2] | **Roi Naveiro**[2,3]

[1]U.S. Air Force Academy, Colorado, Colorado Springs, USA

[2]ICMAT-CSIC, C. Nicolás Cabrera Madrid, Madrid, Spain

[3]CUNEF Universidad, C. de Leonardo Prieto Castro Madrid, Madrid, Spain

Correspondence
Roi Naveiro, CUNEF Universidad, Spain.
Email: roi.naveiro@cunef.edu;

David Ríos Insua (ICMAT-CSIC).
Email: david.rios@icmat.es

Funding information
Air Force Office of Scientific Research, Grant/Award Numbers: 21RT0867, FA-9550-21-1-0239; AXA Research Fund; European Office of Aerospace Research and Development, Grant/Award Number: FA8655-21-1-7042; Fundación BBVA; Horizon 2020 Framework Programme, Grant/Award Numbers: 815003, 101021797; Ministerio de Ciencia y Tecnología, Grant/Award Number: PID2021-124662OB-I00; National Science Foundation, Grant/Award Number: DMS-1638521

Abstract

Automated driving systems are rapidly developing. However, numerous open problems remain to be resolved to ensure this technology progresses before its widespread adoption. A large subset of these problems are, or can be framed as, statistical decision problems. Therefore, we present herein several important statistical challenges that emerge when designing and operating automated driving systems. In particular, we focus on those that relate to request-to-intervene decisions, ethical decision support, operations in heterogeneous traffic, and algorithmic robustification. For each of these problems, earlier solution approaches are reviewed and alternative solutions are provided with accompanying empirical testing. We also highlight open avenues of inquiry for which applied statistical investigation can help ensure the maturation of automated driving systems. In so doing, we showcase the relevance of statistical research and practice within the context of this revolutionary technology.

KEYWORDS

adversarial machine learning, automated driving systems, Bayesian analysis, ethical decision support, heterogeneous traffic, request-to-Intervene

1 | INTRODUCTION

The capabilities of automated driving systems (ADSs) have grown tremendously over the last decade. Recent breakthroughs in artificial intelligence (AI), as well as complementary advances in computational hardware, have allowed cutting-edge perception and control algorithms to be executed in real time. These facts, combined with automobile-electrification trends and evolving views toward vehicle ownership, suggest that a paradigm shift in societal transportation is at hand. An idyllic future characterized by fewer roadway accidents, less pollution, reduced travel times, and increased mobility opportunities (e.g., for the handicapped and elderly) is conceivable.[1] But such a future must be forged. Therefore, within this manuscript, stemming from our recent work in the area, we present a sample of statistical decision-making challenges that must be solved for ADS technology to reach its full potential. By discussing these issues, along with early attempts at their resolution, we endeavor to increase awareness and interest of these topics within the statistical community.

Statisticians have long supported the transportation sector, and their continued engagement is paramount to the development of ADS technology. Among other problem areas, statistical methods have played a major role in the design of

Disclaimer: The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government. Distribution unlimited via PA# USAFA-DF-2022-655.

cost-effective public transportation systems, the refinement of automobile safety, the management of related public health concerns, and the optimization of traffic flows.[2] Akin to these historical achievements, ADS operations entail decision making under multitudinous sources of uncertainty, thereby implying the persistent relevance of statistical techniques in future transportation research. Despite this conceptual similarity, since many relevant sources of uncertainty are unique to the ADS setting, they present myriad, novel issues for the statistical profession.

Many of these problems are, in one manner or another, related to vehicular independence. As an ADS is required to autonomously assess its environment, technological limitations in perception-related sensors are an especially salient source of uncertainty. Precipitation is a known impediment to both visual-light and lidar cameras that introduces errors into their associated data streams. Such environmental effects further complicate scene perception, potentially poisoning the associated image classifier and disrupting subsequent control algorithms. Furthermore, an ADS's reliance on software is another relevant source of uncertainty associated with numerous physical and cybersecurity challenges, for example, data manipulation, privacy vulnerabilities, and malware.[3] Moreover, given that humans cohabitate an ADS's environment, their presence and behavior must be appraised, especially when considering an ADS's interactions with other drivers, pedestrians, cyclists and vehicles. Human behavior is an especially difficult source of uncertainty to manage because of its nonstationarity. Humans are learning agents, and an ADS's presence may influence their behavior.

Whereas the aforementioned uncertainties are only a subset of those associated with ADS operations, they highlight the important challenges that must be resolved before widespread adoption of ADS technology can occur.[4] Qualitatively speaking, the difficulty of these challenges derives from some form of underlying uncertainty implying that, from a methodological perspective, statistical techniques will prove decisive in this domain. Indeed, ADS management offers the statistical community numerous high-impact research problems spanning multiple sub-disciplines. We discuss a subset of these problems herein chosen from our recent work in the ADS domain.

After providing relevant, contextual background in Section 2 for readers unfamiliar with ADS technology or testing, Sections 3 through 6 respectively discuss request-to-intervene (RtI) decision support, ADS ethical decisions, ADS lane-changing in heterogeneous traffic, and the robustification of ADS perception algorithms against adversarial attacks. For each of these sections, we outline the relevant statistical challenge, review earlier research upon it, present solutions and simulation experiments within our previous work, and summarize several open challenges to be addressed in future research. Section 7 provides closing remarks, summarizing the discussions provided herein, and highlights additional statistical challenges in the ADS domain.

2 | CONTEXTUAL BACKGROUND

ADS research is, by its very nature, a multi-disciplinary endeavor. From a technological perspective, ADSs rely mainly on tools and techniques from the fields of statistics, electrical engineering, robotics, computer science, software engineering, and mechanical engineering. Moreover, given the revolutionary effects ADS technology may have on modern society, less quantitative fields of study (e.g., philosophy, sociology, and psychology) are also relevant. Conducting ADS research requires a statistician to understand the current state of knowledge from a multi-disciplinary perspective. Therefore, within this section, we introduce contextual background of this nature and provide references for more in-depth explorations. The topics discussed within this section are foundational to the comprehension of Sections 3 through 6.

2.1 | Current and future state of ADS technology

Incipient ADS technology is beginning to become commercially available. This gradual evolution from manned vehicles (MVs) to ADSs is widely expected to continue over the next twenty years. Vehicles will progressively ascend the six automation level taxonomy* set forth in the Society of Automotive Engineer's (SAE). This taxonomy begins at level 0 (i.e., vehicles having no automated capabilities) and proceeds from levels 1 through 4, such that each level is characterized by increasing autonomous faculties. It culminates at level 5 (i.e., fully automated ADSs with no environmental restrictions).[6]

Prior to 2010, global roadways were populated, almost exclusively, by level-0 automobiles. However, over the last decade, manufacturers have begun to commercially offer vehicles of higher automation levels. For example, *lane keeping assistance* and *adaptive cruise control* systems that, respectively, automatically select steering and acceleration inputs

*Based on reviewer feedback, we note that the SAE taxonomy leaves room for interpretation, for example, whether a driver is present or remote in a level-4 ADS. We adopt the perspective of BMW[5] wherein a driver is present and the ADS has a cockpit.

are now available in vehicles produced by nearly every automotive manufacturer.[7,8] As a consequence, the population of level-1 and level-2 ADSs having limited self-driving features increases annually. Although the majority of ADSs currently available for purchase are level-1, multiple level-2 systems (e.g., Tesla Autopilot, Cadillac Super Cruise) are entering the market. Volvo is dedicated to launching its Ride Pilot level-3 system in 2023. This system is characterized by conditional self-driving automation, which allows the ADS to maintain full control of critical safety functions under specific circumstances. However, the driver is still expected to be aware of the surroundings and be ready to take control of the vehicle in adverse conditions. Similarly, Mercedes-Benz has achieved a remarkable milestone by becoming the first automaker to receive international regulatory approval for producing level-3 vehicles through its Drive Pilot system. Although the Drive Pilot system is currently limited to geofenced portions of highways with a maximum speed of 60 km/h, several other companies, for example, Waymo and Cruise, are testing level-4 (i.e., fully autonomous ADSs within a prescribed operational domain) and level-5 ADSs in real-world scenarios.

Depending upon the problem, different ADS levels are more appropriate to consider than others; however, given the anticipated, gradual progression to fully autonomous ADSs, each tier of technology is a fruitful area of study that offers unique challenges. This is evidenced by the discussion and analyses we provide within this manuscript; we specify in each section the relevant ADS levels addressed.

2.2 | Societal impact of ADS technology

Transportation pervades all aspects of human life and, as a result, changes to its character are far-reaching. Therefore, since the adoption of ADS technology would revolutionize transportation, cascading effects to other aspects of human life should be expected as well. Alternatively, given the interrelationships between transportation and other societal factors, the converse also holds. Myriad exogenous factors also affect society's willingness to adopt ADS technology.

Figure 1, adapted from Caballero et al.,[9] integrates anticipated impacts of ADS technology on society with factors affecting its adoption. This network illustrates the interconnected nature of numerous societal considerations with ADS adoption. In accordance with commonly expressed views in the ADS literature, dotted nodes refer to societal opportunities enabled by ADSs, light gray nodes refer to challenges, and white nodes refer to neutral impacts. Dark gray nodes reference contextual factors (e.g., trust in ADS) that may have a major influence on the massive deployment of ADS.

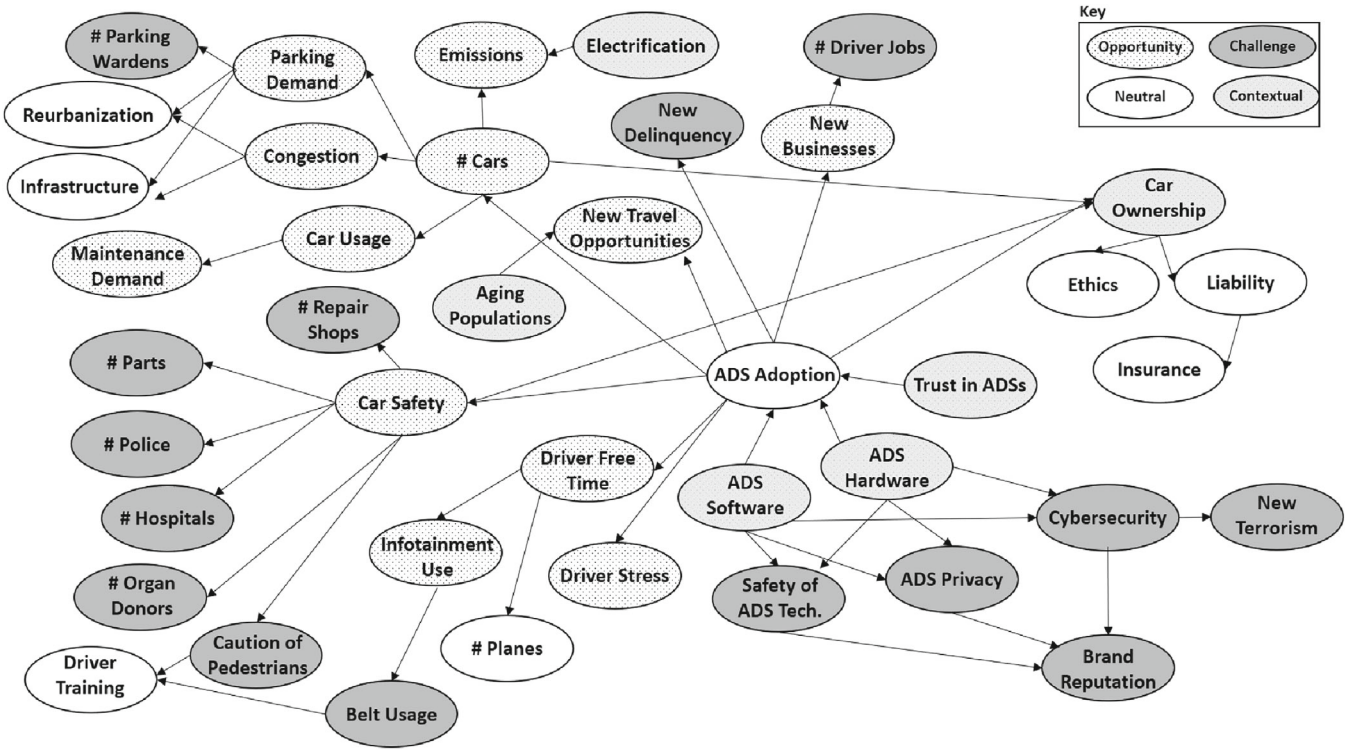

FIGURE 1 Network reflecting ADS impacts on society.

The network emphasizes the multi-disciplinary character of ADS research: whereas technical considerations are of paramount importance, qualitative factors play a significant role as well. This is illustrated in Section 4 via our treatment of ADS ethics and liability within a statistical-decision-making framework.

2.3 | Architectures for ADS decision-making

Alternative frameworks that delineate between distinct layers of an ADS's decision-making architecture are available.[10] Nevertheless, common elements are pervasive. Requisite ADS decision tasks, each of a clear statistical nature, can be summarized as follows:

- Sensory input must be first processed through object detection and localization components to generate scene understanding. The ADS must estimate the scene semantics, understand its geometry, and estimate the position of the vehicle itself.
- The scene understanding must be inputted into a scene prediction component. Obstacles need to be identified and the behavior of other agents in the driving scene must be anticipated.
- Decision components transform scene understanding and prediction into control outputs.

The object detection and localization components are foundational. Often large-scale datasets of varied driving scenes are leveraged therein; several recent approaches[11] have obtained state-of-the-art results with powerful deep learning (DL) techniques. However, even if a driving scene may be precisely perceived,[12] its utility is limited by the accuracy of the scene prediction component. This task is particularly challenging given that many obstacles are dynamic learning agents (e.g., pedestrians). It is also critically important given its connection to the decision components. Whereas some of these decision components are less dire (e.g., route planning), others are associated with potentially fatal consequences (e.g., motion planning or request-to-intervene decisions).

Techniques for perceiving and forecasting the environment are often borrowed from other application domains (e.g., image classification in social networks). Varieties of convolutional neural networks (CNNs), as well as state-space models, are commonly leveraged. Alternatively, whereas ADS control algorithms were originally characterized by collections of conditional, environmental rules (e.g., based on the presence of lane markings and guard rails), more sophisticated techniques have since been developed. Such methods often adopt a DL framework to automatically learn relevant features from training examples.[13] These techniques may determine steering and speed controls given input images of the road ahead; in this manner, sensor inputs are linked to control outputs, following an end-to-end approach. Methodologically speaking, components are increasingly implemented via jointly trained DL algorithms. This is an attractive paradigm because it allows each component to form an optimal representation for the desired end task. The PilotNet architecture[12] is an exemplar of such a system that performs autonomous control without distinct internal modules connected by human-designed interfaces. Deep reinforcement learning methods can also serve to seamlessly streamline ADS decision-making, as described by Grigorescu et al.[13]

Nevertheless, although significant strides have been made in the design of ADS architectures, numerous statistical challenges remain, especially when viewed from a statistical decision-theoretic perspective. Such problems are often holistic in nature, pertaining to the operationalization of statistical insights into ADS decisions.

2.4 | Testing ADS technology

Given that ADSs must interact with humans, there always exists the risk of a fatal accident.[14] This risk is exceptionally elevated when experimenting with novel ADS architectures and algorithms. Therefore, ADS testing occurs upon a ladder in which environmental realism is progressively increased. Experiments typically start on a relatively simplified numerical simulator; if results are satisfactory, the novel ADS system is implemented in a more realistic computational simulator. Subsequently, a three-dimensional virtual test environment is used, potentially augmented with human agents. Experiments with a physical ADS in a closed environment follow. Finally, testing culminates with experimentation being performed within a controlled real-world environment.

This iterative experimental design is accompanied by its own statistical difficulties (e.g., the optimal combination of design points across related but disparate conditions). The design of the simulators themselves (e.g., which uncertainties

are modeled and how) is also a critical statistical challenge that determines the information garnered via experimentation. Although we introduce such challenges herein, we set aside their detailed consideration for future research.

Within this research, we focus on simpler, numerical simulators to illustrate potential decision making frameworks. Whereas the simulated environment used herein is exclusively for demonstration purposes, it is typical of the first stage of ADS-life-cycle testing. It is is implemented as a discretized roadway composed of equal-length cells. The configuration of the roadway, ADS, driver-state definitions, environment, trajectory planning module, and ADS's preference model is determined by a user-defined set of hyperparameters. These hyperparameters can specify, for example, the roadway's length in terms of the number of cells, the ADS's operational design domain (ODD), the driver's awareness level, the types of obstacles present in the environment, the driver-and-environment transition functions and so forth. Once these hyperparameters are defined, their settings can be adjusted for each design point to reflect the corresponding experimental conditions. For instance, the hyperparameters can be modified to reflect variations in trip duration or the likelihood of encountering specific obstacle types. Our environment is implemented as a class in Python and is available for use or modification[†].

Our simulator employs five primary methods. The first method, called *predict*, generates forecasts for the environment and driver state at each position of the autonomous driving system (ADS). The second method, *update*, uses newly observed information to estimate the current driver and environment states. The third method, *decide*, determines the control outputs for the current position and plans a trajectory for subsequent ones. Additionally, the system includes an *issue warnings* method that evaluates forecasts and emits warnings if necessary when using level-3 and level-4 ADS architectures. The fourth method, *evaluate driving modes*, decides whether human or ADS control is preferable at any given time. Finally, the fifth method, *move*, synchronizes the other methods. When called, the ADS advances one cell, the *update* method processes new information, the *predict* method generates relevant forecasts, and the *decide* and *issue warnings* methods determine appropriate control outputs based on these forecasts.

The forthcoming sections propose novel solutions to several relevant architectural design decisions in ADS and illustrate the use of our simulation environment to test the soundness of the proposed concepts prior to more detailed tests in a realistic computational simulator further up in the testing ladder.

3 | THE REQUEST-TO-INTERVENE DECISION

In level-3 and level-4 ADSs, given the potential consequences of a failed control transfer, the precise manner in which control transfer occurs is of paramount importance. One promising methodology entails the ADS asking the driver to take control via a *Request-to-Intervene* (RtI) command. Such a command would be executed when ODD conditions are approaching their limits or have been exceeded.[15] Until level-5 vehicles predominantly populate global roadways, the management of RtIs will remain a crucial, safety-related issue. The remainder of this section sketches a statistics-based approach for managing RtI decisions, presents initial empirical results, and discusses open problems. For a more detailed examination of this problem and the methods discussed herein, we refer an interested reader to Ríos Insua et al.[16]

3.1 | A predictive methodology for RtI management

Consider a level-3 or -4 ADS for which three driving modes are available: *automated*, *manual*, and *emergency*. Given the ODD restrictions of such a vehicle, RtIs must be algorithmically managed in a continuous fashion. However, the uncertain temporal evolution of the underlying system state[‡] is a complicating factor; the ability of the ADS to perceive the system state and forecast its evolution considerably affects its decision quality.

Therefore, we present a framework for RtI management that aggregates historical information for inference and prediction. This framework is underpinned by the qualitative process depicted in Figure 2. A Bayesian approach that iteratively leverages incoming data is utilized to update beliefs about the system state and forecast its evolution. D_t designates the compendium of data collected up to time t (e.g., sensor input). In application, ADS decisions are often made considering a rolling time horizon (e.g., $k = 10$ discretized time intervals of 0.5 s each); a similar approach is thus adopted herein.

[†]Two distinct implementations of this environment are available at https://github.com/roinaveiro/ads_trusto and https://github.com/roinaveiro/preferences_ads, respectively.

[‡]We use the term system state to represent the collective state space that incorporates both the driver and environmental states.

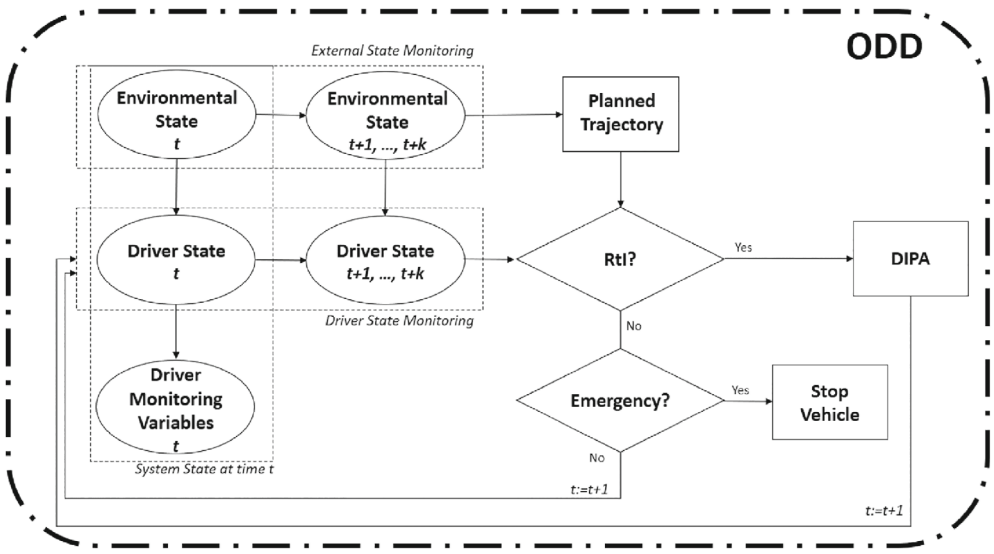

FIGURE 2 Framework for RtI inference, prediction and decision support.

Although the number of intervals may be based on the current system state, we assume, without loss of generality, that it is a fixed parameter.

The framework depicted in Figure 2 emphasizes the *management by exception principle*.[17] A group of models is used for inference, prediction and decision support under standard driving conditions until an exception (i.e., a likely ODD violation) arises that triggers an intervention request. This RtI decision-making process can be summarized as follows. The environmental and driver state monitoring (DSM) systems observe the state at time t, forecast future states through time $t + k$, update the planned trajectory, decide whether an RtI should be executed[§] and, if so, evaluate such execution via a driver intervention performance assessment (DIPA). This iterative process is then repeated at time $t + 1$. Each of the aforementioned components, and how they affect the RtI management decision are discussed in turn.

3.1.1 | Operational design domain supervision

Since an ADS must operate within its prescribed ODD,[15] conditions under which the automated mode may be utilized are constrained. The ODD typically is characterized by the roadway state (e.g., road or visibility conditions), the behavior of the ADS (e.g., speed), and the state of the vehicle's physical subsystems (e.g., tire pressure). Therefore, we define $\boldsymbol{g}_t^1$, $\boldsymbol{g}_t^2$, $\boldsymbol{g}_t^3$ as distinct blocks of variables for each respective category. These variables are checked at time t and concatenated into a single vector defined as $\boldsymbol{g}_t = (\boldsymbol{g}_t^1, \boldsymbol{g}_t^2, \boldsymbol{g}_t^3) \in \mathcal{G}$ wherein $\mathcal{G}$ represents the ODD under which the ADS may operate autonomously.

Although the evolution of some of the variables can be accommodated with deterministic physical models, to maintain generality and account for observation error, we utilize probabilistic models to forecast all ODD variables. A flexible strategy to do so employs state space models.[17] Given the short planning horizons typically considered, an important example is the local level (i.e., random walk plus noise) model where $\boldsymbol{G}_t = \boldsymbol{F}_t + \boldsymbol{V}_t$, and $\boldsymbol{F}_t = \boldsymbol{F}_{t-1} + \boldsymbol{W}_t$, where $\boldsymbol{G}_t$ are observable ODD variables (with realizations $\boldsymbol{g}_t$); $\boldsymbol{F}_t$ are unobservable state variables; and, $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ are vectors of random noises. With this information, a k-period forecasting model $p(\boldsymbol{G}_{t+k}|D_t)$ can be recursively constructed for the ODD variables. Based on this forecasting model, queries concerning $Pr(\boldsymbol{G}_{t+k} \notin \mathcal{G}|D_t)$ may be addressed. Should this probability be sufficiently large, the ADS may determine that ODD limits are likely to be exceeded, implying that an alert should be issued and the automated driving mode abandoned. Moreover, $p(\boldsymbol{G}_{t+k}|D_t)$ may also be used to issue warnings when unlikely changes in the operational conditions are detected. Even if such changes occur within ODD limits, they may be indicative of rapidly changing operational conditions requiring increasing driver awareness, but no intervention.

[§]Note that, if a contingency arises and the ADS assesses the driver is incapacitated, an emergency maneuver will be executed autonomously.

3.1.2 | Environment monitoring

This pertains to the prediction of highly dynamic conditions external to the ADS but that substantially affect its operations. Without loss of generality, assume there are η of such variables captured by $\boldsymbol{Y}_t = (Y_t^1, Y_t^2, \dots, Y_t^\eta)$. As tangible examples, these may correspond to the behavior of objects, persons, or other vehicles in the driving scene. More stable roadway environment conditions (e.g., weather) will be typically included within the ODD $\boldsymbol{g}_1^t$ variables.[15]

Akin to the ODD variables, a forecast for the environmental variables between two consecutive time steps, that is, $p(\boldsymbol{Y}_{t+1}|\boldsymbol{Y}_t)$, may be constructed using a state space model. From these quantities the k-steps forecast may be calculated via the recursion $p(\boldsymbol{Y}_{t+i}|D_t) = \int p(\boldsymbol{Y}_{t+i}|\boldsymbol{Y}_{t+i-1})p(\boldsymbol{Y}_{t+i-1}|D_t)d\boldsymbol{Y}_{t+i-1}, i = 1, 2, \dots, k$. As previously discussed, this forecasting model would also be utilized for driving-mode selection and warning-issuance support.

3.1.3 | Driver state monitoring

When an RtI is issued and the driver is unattentive, there is an elevated risk of failure. Therefore, DSM is an essential component of RtI decisions in level-3 and -4 ADSs; the decision support system must understand the awareness of the driver to determine their ability to assume control of the vehicle. Much like the monitoring of the external environment, a DSM system includes a variety of sensors used to perceive the driver's state variables (e.g., fatigue and distraction).

The driver state is modeled by a latent variable $\boldsymbol{\theta}_t \in \Theta$; at each time step t, the ADS collects n sensor measurements $\boldsymbol{X}_t = (X_t^1, \dots, X_t^n)$ from the driver. These sensor measurements are assumed to be related to the driver's true state (e.g., the posture of a tired driver). Moreover, the environment monitoring variables, $\boldsymbol{Y}_t$, are assumed to affect the driver's state (e.g., increased attention during a storm). To account for these dependencies, we derive a forecasting model, $p(\boldsymbol{\theta}_{t+i}|D_t)$, for some horizon $i \in \{1, 2, \dots, k\}$, from the following conditional distributions.

Namely, in addition to the environment-monitoring forecast, that is, $p(\boldsymbol{Y}_{t+1}|\boldsymbol{Y}_t)$, the DSM forecasting model is specified by $p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, \boldsymbol{Y}_{t+1})$, $p(\boldsymbol{X}_{t+1}|\boldsymbol{\theta}_{t+1}, \boldsymbol{X}_t)$, $p(\boldsymbol{\theta}_1|\boldsymbol{Y}_1)$, $p(\boldsymbol{X}_1|\boldsymbol{\theta}_1)$, and $p(\boldsymbol{Y}_1)$. The predictive evolution of the driver state given their previous state and the current environmental variables is given by $p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, \boldsymbol{Y}_{t+1})$. Alternatively, $p(\boldsymbol{X}_{t+1}|\boldsymbol{\theta}_{t+1}, \boldsymbol{X}_t)$, describes the predictive evolution of the sensor measurements given such measurements at time t and the driver state at time $t+1$. Finally, $p(\boldsymbol{\theta}_1|\boldsymbol{Y}_1)$, $p(\boldsymbol{X}_1|\boldsymbol{\theta}_1)$, and $p(\boldsymbol{Y}_1)$ respectively represent the distribution over driver states given the initial environmental variables, the distribution over the sensor measurements given the initial driver state, and the distribution over the environmental variables. Based on these, $p(\boldsymbol{Y}_{t+1}|\boldsymbol{Y}_t)$ can be derived via the recursive computations in Ríos Insua et al.;[16] the resultant probabilities may then be used to inform driving-mode and warning-issuance decisions.

3.1.4 | Trajectory planning

ADSs are equipped with trajectory planning systems that guide their movements.[18] At a given time t, these systems provide a trajectory plan $\overline{\boldsymbol{z}} = \{\boldsymbol{z}_{t+1}, \boldsymbol{z}_{t+2}, \dots, \boldsymbol{z}_{t+k}\}$ over the required time k that attempts to maintain the ADS within its ODD limits. In turn, this planned trajectory is utilized as input for RtI management decisions.

3.1.5 | Driving mode assessment

Provided some trajectory plan $\overline{\boldsymbol{z}}$ at time t, consider assessing a driving mode d over a length-k horizon. The utility of this driving mode depends upon the environmental state, the driver state and the status of the relevant ODD variables. The evolution of the environmental and driver states are unknown, but their values may be forecast. Let $\overline{\boldsymbol{Y}} = [\boldsymbol{Y}_{t+1}, \dots, \boldsymbol{Y}_{t+k}]$ represent a predicted sequence of environmental states and $\overline{\boldsymbol{\theta}} = [\boldsymbol{\theta}^{t+1}, \dots, \boldsymbol{\theta}^{t+k}]$ capture a predicted sequence of driver states. The joint probability of these sequences occurring can be calculated via

$$p(\overline{\boldsymbol{Y}}, \overline{\boldsymbol{\theta}}|D_t) = p(\boldsymbol{Y}_{t+1}|\boldsymbol{Y}_t)p(\boldsymbol{\theta}^{t+1}|\boldsymbol{Y}_{t+1}, D_t)\prod_{i=2}^{k} p(\boldsymbol{\theta}^{t+i}|\boldsymbol{\theta}^{t+i-1}, \boldsymbol{Y}_{t+i}, D_t)p(\boldsymbol{Y}_{t+i}|\boldsymbol{Y}_{t+i-1}).$$

Furthermore, a utility function $u(d, \overline{\boldsymbol{z}}, \overline{\boldsymbol{Y}}, \overline{\boldsymbol{\theta}}, \boldsymbol{g}_t)$ can be defined to assess the efficacy of the driving mode over the next k steps, incorporating the ADS's objectives with information from time t and forecasts for times $t+1$ to $t+k$. Such a

Algorithm 1. ADS controller

Input: Utility function; priors over the ODD, environment, and driver-state variables.
Output: Trajectory from ORIGIN to DESTINATION
(and implementation of commands when in AUTON or EMERG modes).
while DESTINY not reached **do**
  Read internal and external sensors.
  Forecast Environment k steps ahead.
  Forecast driver state k steps ahead.
  Compute trajectory.
  Assess and select DRIVING MODE. If necessary, WARNINGS.
  Manage from DRIVING MODE. Resolve any pending DIPAs.
end while

utility function thus takes into account, the driving mode, trajectory, environment forecast, driver-state forecast, and ODD conditions. For the present, we assume that this utility function is exogenously defined; however, a framework for its development is described in Section 4.

Since the future system state is a random variable, we compute the predictive expected utility of the driving mode via $\psi(d) = \int\int u(d, \overline{\mathbf{z}}, \overline{\mathbf{Y}}, \overline{\boldsymbol{\theta}}, \mathbf{g}_t)\, p(\overline{\mathbf{Y}}, \overline{\boldsymbol{\theta}}|D_t)\, d\overline{\mathbf{Y}}\, d\overline{\boldsymbol{\theta}}$. The expected utility of driving mode d is foundational for our driving mode assessment and selection. By leveraging its value, a ranking over driving modes can be constructed to decide which is most appropriate.

3.1.6 | Driver intervention performance assessment

Within RtI management, it is helpful to understand how capable a driver is of intervening. DIPAs can be used to learn this by recording and retaining historical, driver-intervention data.[19] Namely, after an RtI has been issued, the ADS can compare the driver's actual and hypothesized performances. This information may then be used to update beliefs used in future RtI decisions.

Assuming an RtI has been executed over $[t, t+k]$, the driver's performance is evaluated via some observed $u(d_1)$, wherein d_1 designates the manual driving mode. After completion of the RtI, $u(d_1)$ is compared to the driver's expected performance, that is, $\psi(d_1)$. Our approach to DIPAs is characterized by identifying whether a driver under- or over-performed; if $\psi(d_1) - u(d_1) > 0$ the driver underperformed, otherwise they overperformed. Inference regarding the underperformance probability q can be accomplished using a beta-binomial model.[20] This value may be leveraged to modulate the ADS's risk aversion.

3.1.7 | Driving mode transitions

Having discussed the foundational elements of our RtI management framework, we describe how they may be jointly leveraged to manage transition between driving modes. While operating in automated mode, the ADS should periodically issue predictive, ODD-compliance risk assessments. If $Pr(\mathbf{g}_{t+k} \notin \mathcal{G}|D_t)$ is sufficiently great, the ADS should alert the driver, and assesses whether the automated or manual driving mode (i.e., d_0 or d_1) is preferable. Based on this assessment, an RtI may or may not be issued to the driver. If issued, a DIPA should be performed following the intervention. If the driver performs too poorly, the ADS may assess that the driver is incapacitated and trigger the emergency mode (i.e., d_2). Otherwise, the driver retains control until they desire to return to automated mode. Conversely, if while calculating $Pr(\mathbf{g}_{t+k} \notin \mathcal{G}|D_t)$, the ADS determines ODD limits are being critically approached, no RtI should be issued; instead, the emergency mode should be undertaken immediately and the corresponding alert issued. This emergency mode should bring the ADS to a stop safely and, after this point, the driver may select a driving mode upon resuming operations.

This ADS management framework is qualitatively and holistically summarized in Algorithm 1. Detailed subroutines for the processes described previously are provided in Ríos Insua et al.[16]

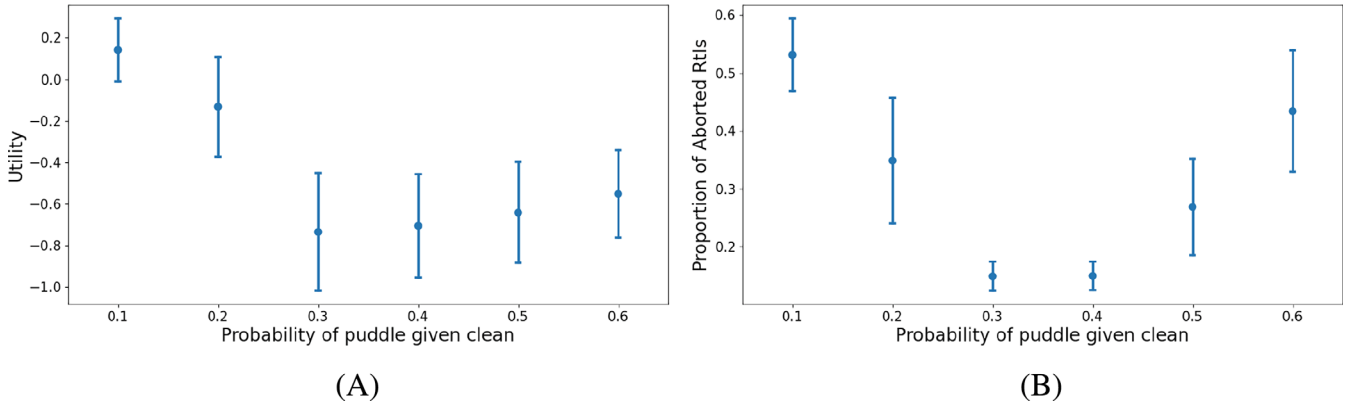

FIGURE 3 Performance measures for varying levels of $p(y_{t+1} = \text{puddle} | y_t = \text{clean})$. (A) Attained utility. (B) Proportion of aborted RtIs.

3.2 | Empirical evaluation of the RtI framework

Following the testing lifecycle sketched in Section 2.4, prior to testing in more realistic simulation environments, the proposed framework was thoroughly tested using our numerical simulator. The focus of the testing was to simulate complex driving scenarios with dangerous conditions in order to ensure that the framework is robust enough to handle such situations. To achieve this, we simulated an ADS's trip along a 90-cell road multiple times; each cell is described as being clean or containing one of two types of obstacles (i.e., puddles or rocks). The ADS was equipped with the driving mode management strategy presented in Algorithm 1. Several performance-related measures (e.g., the total attained utility or the number of crashes) were calculated to assess the strategy's effectiveness. Detailed results of these experiments can be found in Ríos Insua et al.,[16] which demonstrate that the proposed strategy consistently performs in a reasonable manner.

Of particular interest is the fact that the experimentation highlights a phenomena referred to as the *fundamental dilemma* for level-3 and level-4 ADS. That is, when ODD limits are likely to be exceeded, the driver may not be prepared to assume control of the vehicle. In such situations, the ADS must decide whether to transfer control to a distracted driver. But, in so doing, the ADS may be forced to make life-or-death decisions. When considering such situations during ADS design, it is unclear which course of action is preferable. If control is transferred to the distracted driver, the human is forced to assume responsibility for their distraction. Conversely, if the ADS retains control, a better outcome may be obtained but at the expense of automating portentous decision-making. Such questions do not have simple answers, but we illustrate a potential resolution based on maximizing expected utility.

The behavior of the graphs in Figure 3 typifies how our approach resolves the *fundamental dilemma*. Dangerous roadway conditions are characterized by slick pavement; the conditional probabilities presented refer to the presence of a puddle in the roadway given no other obstacles. Figure 3A depicts the total attained utility achieved by the ADS that follows the driving mode management strategy proposed. Figure 3B represents the proportion of aborted RtIs. An aborted RtI occurs when the ADS predicts with high probability that the ODD limits will be exceeded, but determines that the automated driving mode yields greater expected utility than the manual mode. The proportion of aborted RtIs is calculated as the fraction of all RtIs that were aborted. Estimates are computed across several simulations. Mean plus-minus two standard deviations are depicted. When increasing the probability of puddles beyond a certain limit, we observe that the utility starts to increase in Figure 3A. The reason for this phenomenon is that, when the roadway is anticipated to be especially hazardous, the ADS more frequently detects driver under-performance. In turn, via our expected utility maximization framework, the ADS decides not to allow the driver to take control over the vehicle as often. This behavior can be observed in Figure 3B wherein above a certain value of the probability of puddles, aborted RtIs start increasing and thus the risk of having a distracted driver in control is diminished which, in turn, increases utility.

3.3 | Open RtI management problems

Whereas the framework discussed herein proved successful in our simple numerical simulator, to more thoroughly validate this approach, additional testing is required within higher-fidelity settings. However, this added realism presents

challenges. For example, our framework relies upon a host of conditional distributions and, whereas they may be defined via subjective Bayesian priors, accuracy is likely to increase as these distributions are learned from real data. This implies that empirical studies on the evolution of the environmental and driver state, the relationship between the environment and drive state, and the relationship between the driver state and sensor measurements are a critical need.

From an algorithmic perspective, a more systematic method for identifying driving-mode and warning-issuance thresholds is required. The illustrations set forth by Ríos Insua et al.[16] leveraged a simple search algorithm to identify these parameters; however, such an approach is likely to prove itself infeasible for more realistic environments. Given the degree to which these thresholds affect ADS performance, their proper identification is imperative. Future research leveraging approximate dynamic programming to this end is particularly promising.

Finally, because our RtI management framework is rooted in expected-utility maximization, the definition of the underlying utility function is foundational. The pioneering work of Awad et al.[21] provide a structure upon which such functions may be created. In the next section, we describe this problem in more detail, along with initial work towards its resolution.

4 | ETHICAL ADS DECISION-MAKING

As is often the case with revolutionary technologies, the widespread adoption of ADSs is accompanied by myriad moral uncertainties, as illustrated, for example, through the vignettes presented by Lin.[22] Unfortunately, in addressing these issues, decision makers are forced to grapple with unenviable ethical predicaments. The fundamental dilemma discussed in Section 3.2 is a prime example. Unitary best responses do not exist to such problems; however, we contend that flexible frameworks can be developed to incorporate diverse, ethical perspectives.

This contention is implicit within our approach to the fundamental dilemma of Section 3. By addressing ADS control with statistical decision theory, the selected utility function forms the basis of the ADS's behavior and dictates the ethos underpinning its choices: by directing the ADS's choices, the utility function implicitly defines the morality of its behavior. Therefore, the purposeful design of these functions is of paramount importance; if one desires the ADS to adopt a particular ethos over another, they need only embed this perspective within the utility function.[23]

In order to enable meaningful analyses and comparisons, stakeholders (e.g., manufacturers, regulators) must be able to tailor an ADS's utility function to incorporate diverse operational and regulatory considerations as well as ethical perspectives. By doing so, uncertainty about the ADS's preferences can be reduced, and thoughtful design and regulation can be achieved, with these developments clearly affecting level-3 to -5 ADSs. Moreover, standardizing the ADS' decision making through an expected-utility-maximization approach makes this process fully transparent and reproducible, a beneficial feature for myriad applications (e.g., determining accident liability).

In the remainder of this section, we discuss a general, decision-analytic approach toward utility function design that allows for the purposeful selection of an ADS's ethical framework and operational priorities. Initial empirical results and open areas of inquiry are provided as well.

4.1 | Ethical ADS decisions

The relative importance of the objectives pursued by an ADS are stakeholder dependent. A private owner may seek to maximize comfort and minimize trip duration. An insurance agency is likely to be more concerned with vehicular safety than driver comfort. A freight-shipping company may be interested in maximizing profit while preserving corporate reputation. Two insurance agencies having distinct ethical perspectives may evaluate vehicular safety via different weights. To accommodate these diverse perspectives, we developed a generic multi-attribute utility model for ADS management that allows designers, owners and policymakers to tailor ADS behavior according to their own priorities and ethical framework.[24] This requires the identification of an encompassing set of objectives along with a collection of attributes for their assessment. Analysis was also provisioned regarding how selection among these objective-attribute combinations, as well as the utility function's structure, affect the ADS's ethos. Akin to Keeney,[23] our framework can accommodate multiple

TABLE 1 Summary of objectives and attributes.

Upper level obj.	Lower level obj.	Natural attribute	Constructed attribute[a]	Proxy attribute
Performance	Min. fuel consumption	Monetary units		
	Min. trip duration	Monetary/temporal units		
	Min. passenger discomfort		Yes	ADS movement
Safety	Min. injuries of individuals inside (outside) ADS	Number of injuries	Yes	No. in hospital
	Min. fatalities of individuals inside (outside) ADS	Number of fatalities / VSL	Yes	
	Max. respect for life	Probability of death/injury	Yes	
	Min. damage to ADS	Monetary units	Yes	
	Min. infrastructure damage	Monetary units	Yes	
	Min. environmental impact (global/local)	Monetary units emissions	Yes	
Reputation	Min. harm to manufacturer reputation		Yes	Media salience
	Min. harm to societal perceptions		Yes	Media salience

[a]See Caballero et al.[24] for the constructed-attribute scales.

ethical viewpoints; however, we primarily focused on the consequentialist[¶] and deontological[‖] paradigms, as well as a subset of their refinements (e.g., rule utilitarianism).

The derived set of ADS objectives is presented in Table 1. The first column represents the upper-level objectives, whereas the second column sets forth the lower-level objectives.[25] To capture the most commonly considered impacts, these objectives were identified via an in-depth review of the transportation literature with specific emphasis on ADS, multiple-objective, regulatory, ethical, and safety applications. A mind map was used to group the identified objectives, and these were subsequently segmented into a tree having upper- and lower-level objectives, then validated by an external panel of subject-matter experts to ensure compliance with canonical requirements.

These objectives are not useful from a decision-analytic perspective without corresponding attributes. Potential attributes for each lower-level objective are summarized in the last three columns in Table 1; multiple alternatives for several objectives are considered. Generally speaking, it is advisable to leverage natural attributes whenever available. If they are not available or the stakeholder deems them inappropriate (e.g., based on ethical implications), then a constructed scale should be utilized. Alternatively, if the stakeholder perceives the constructed attribute to be too ambiguous, or insufficient for any other reason, then a proxy attribute should be used instead. A detailed presentation of each of these attributes is provided by Caballero et al.,[24] along with discussion regarding their relationship to the properties set forth by Keeney and Gregory.[26]

To be operationalized, a set of objective-attribute pairs need only to be equipped with a preference model structure and objective weights. A traditional, decision-analytic approach would leverage an elicitation procedure to determine these quantities. Such an approach is valid in our setting; however, given the emphasis on ethical decision making, care should be taken in its execution.

The ethics associated with a multi-attribute utility function are determined by (1) the selected objectives, (2) the associated attributes, (3) the preference model's functional form, and (4) the objective weights utilized. Therefore, an ADS which only considers the safety of its passengers is adopting an egoist perspective, whereas an ADS that focuses on pedestrian safety employs more of an altruistic ethos. Moreover, a deontological ethos (i.e., one associated with a moral imperative) may be captured via a constructed attribute that only provides positive utility to the (perceived) ethical action. A consequentialist ethos may be captured via more traditional, natural attributes (e.g., number of injuries). Such ethical consequences must be considered during objective-and-attribute selection.

Similar dynamics hold when selecting a preference model, for example, the functional forms from expected utility theory or cumulative prospect theory.[27] For example, assuming that a multiplicative utility function is leveraged, consequentialist thought can be modeled using any objective weighting. This implies that, from a structural perspective, this

[¶]The ethical perspective asserting that the morality of an action depends upon its consequences. An action has no ethical character in and of itself.
[‖]The ethical framework emphasizing the inherent morality of action. An action is right or wrong in and of itself, regardless of its consequences.

ethos and preference model coincide well. Conversely, this is not necessarily the case from a deontological perspective. If a given stakeholder is characterized by a deontological ethos for some objective and its weight does not equal zero, then the marginal utility of this objective may change given different values of another objective's attribute. This logic is often better suited for consequentialism, implying the multiplicative model may be inappropriate.

Moreover, the selection of objective weights brings in an additional layer of nuance. Notably, from a consequentialist perspective, the ratio of the objective weights can be used to partially infer the ADS's ethical perspective. If a much higher weight is assigned to protecting the ADS passengers compared to individuals outside of the vehicle, then an egoist ethos is being adopted. Should the converse hold, then the ADS is adopting an altruistic ethical perspective. Alternatively, if the weights are approximately equal, an egalitarian ethos is being utilized. Intermediate weights between these values enable nuanced mixtures of ethical frameworks to be constructed as well.

Viewed in this light, the advantages of our generic multi-attribute utility model are apparent. Given that our approach simultaneously addresses operational and ethical concerns, not only can it be used for ADS control, but it can be readily leveraged for regulatory purposes as well (e.g., by mandating a preference model, objective weights, etc.). Such regulation could be enacted by either governmental or corporate entities; in turn, compliance with such regulation can also inform criminal and liability proceedings, among others. Therefore, by explicitly considering the ethos of the ADS's utility function, the framework discussed can be utilized to inform solutions for multitudinous problems across myriad stakeholders.

4.2 | Empirically evaluating the framework

To explore the efficacy of this framework, testing was conducted within our simulated environment to investigate how varied parametric settings affect ADS behavior. We investigate the impact that an ADS's utility function has on multiple performance measures, thereby illustrating the effects of stakeholder tuning. Such an approach may be interpreted as a dual to the famous moral-machine experiments:[21] whereas those experiments presented potential consequences to assess and infer individual preferences, we directly assess preferences and explore their consequences empirically.

A consequentialist-normative perspective was adopted during experiments. Of the objectives presented in Table 1, we considered trip duration, harm to individuals inside the ADS, and harm to individuals outside the ADS. We performed two blocks of experiments. The first illustrates that the framework induces reasonable ADS behavior and establishes a benchmark. It also assesses two trade-offs: trip duration versus safety, and passenger-centric versus altruistic safety (i.e., safety of individuals inside vs. outside the ADS). Alternatively, the second block of experiments showcases how liability determination can be addressed utilizing the developed framework. Detailed analysis of the first block of experiments is provided by Caballero et al.[24] Within this manuscript, our discussion is focused on the second block.

As mentioned, a main advantage of our framework is that it renders transparent the decisions taken during ADS design, production and operations. Regulators can leverage this feature by undertaking in-depth simulations of various configurations until they arrive at socially acceptable results. Such configurations can be mandated by law or recommended as industry standards. Consider liability determination as a tangible example, and assume a regulator has set some safety criteria that must be met by any ADS operating on public infrastructures. The safety criteria do not distinguish between individuals inside and outside the vehicle (i.e., both are equally weighted). These are expressed, for example, as follows: *Mean plus two standard deviations of number of injuries and fatalities per X miles should not be greater than 1.4 and 0.25, respectively*. Based on this, the regulator wishes to determine a recommended industry standard for the ADS objective weights.

By simulating ADS operations, differing objective weights can be analyzed and, for each setting, the regulator can determine whether the safety criteria is met. Figure 4A illustrates that if the inside safety weight is fixed at 0.1, trip duration weights greater than or equal to 0.2 do not meet the regulator's criteria (depicted through the green and yellow dotted lines). With this in mind, trip duration weights below 0.2 could be further explored to identify the maximum weight fulfilling the safety constraints. In this particular case, a trip duration weight of 0.1 is admissible per the safety criteria.

Having completed their analysis, the regulator identifies a weight combination as an industry standard; the weights for inside safety, outside safety, and trip duration are 0.1, 0.1, and 0.8, respectively. However, assume an auto manufacturer determines it can gain market share by allocating more weight to inside safety while maintaining the trip duration weight constant, thereby decreasing the outside safety weight. Concretely, suppose that an inside safety weight of 0.7 is selected. If an injury or fatality occurs, it is natural to consider whether the auto manufacturer is liable due to their deviation from the industry standard. To make this determination, a simulation of the particular weight configuration may be

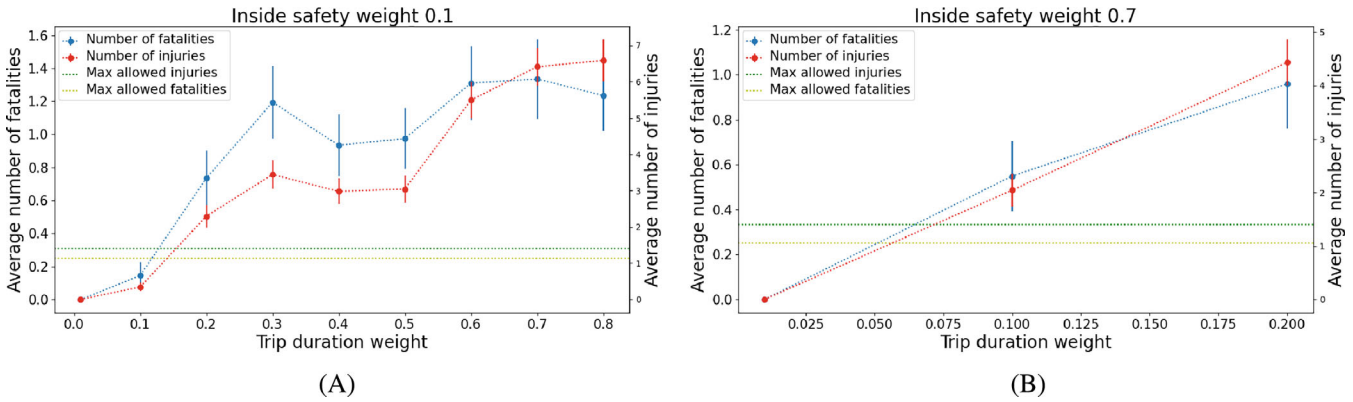

FIGURE 4 Average number of injuries and fatalities versus trip duration weight. (A) Inside safety weight of 0.1. (B) Inside safety weight of 0.7.

undertaken to ascertain whether the results comply with the prescribed safety criteria. Figure 4B depicts that, in this particular example, the manufacturer's selected weights do not meet the regulator's criteria. Therefore, the manufacturer could reasonably be deemed liable; indeed, an inside safety weight of 0.7 reduces the hazard faced by the ADS occupants but with a corresponding increase in risk to pedestrian safety.

4.3 | Open statistical problems in ADS ethics

Few authors expressly examine the ethics associated with a multi-attribute utility model; often a consequentialist ethos is utilized without additional consideration. The augmentation of such inquiry, be it related to ADS applications or other autonomous systems, is an imminent need. Furthermore, although our framework provides a mean upon which ethical ADS behavior can be defined, further exploration is necessary. Future research should use our framework to develop competing ethical models and contrast their behavior in varied traffic conditions. Higher fidelity simulation environments may also be leveraged to inform more realistic regulatory benchmarks. ADSs should be at least as safe as MVs; therefore, real-world benchmarks from federal transportation agencies may suffice as initial baselines. Unfortunately, there is no guarantee that these MV baselines will be societally acceptable for ADSs. The moral-machine experiments illustrated that ADS acceptance may be less rooted in aggregate behavior than in context-dependent settings. As such, future research may combine the empirical approach of the moral-machine experiments with the prescriptive models developed herein to reason about societal preferences in an actionable manner. Inferring objective weights from empirical data for the purpose of learning moral frameworks actually used (not just self-reported) by human decision makers, is a compelling statistical problem as well. Initial attempts along this line of research are reported by Kim et al.[28] Finally, as with other concepts presented in this article, experiments higher up the testing ladder should be undertaken.

5 | ADS OPERATIONS IN HETEROGENEOUS TRAFFIC

Due in part to persistent technological limitations, regulatory dilemmas and consumer inertia,[29] traffic on global roadways will be heterogeneous for many years; MVs will operate alongside ADSs of varying levels, and we are only beginning to discern the associated effects.[30] However, there currently exists a high degree of uncertainty about ADS-to-ADS and ADS-to-MV interactions that affects all manner of operations. Thus, heterogeneous transportation environments are associated with a broad array of statistical problems, especially in relation to ADS decision support.

Concretely, an ADS may interact with traffic in various settings (e.g., car-following, lane-keeping and lane-changing). Each situation is characterized by its own subtleties. As such, a melange of solution approaches have been leveraged to provide decision support. For example, assuming homogeneous traffic of level-5 ADSs, Hult et al.[31] illustrate how such vehicles can cooperate at an intersection with computations implemented via mathematical programming. However,

when cooperative behavior cannot be assured, alternative techniques have been leveraged. Notably, game-theoretic reasoning has frequently been used to analyze aggregate behavior of multi-agent roadway systems. While tractable, these game-theoretic approaches are often sensitive to common critiques of the field (e.g., the perfect-rationality or common-knowledge assumptions). Other authors have leveraged methods from control theory,[32] multi-objective optimization,[33] and cellular automata models[34] in similar situations to curtail such criticism. Di and Shi[35] and Schwarting et al.[36] provide excellent surveys summarizing such developments, thereby illustrating the relative dearth of research focused on the heterogeneous setting. Nevertheless, it is precisely this hybrid, human-machine decision space (i.e., *mixed autonomy*) that currently deserves concerted research.

Regardless of the solution methodology adopted, because our collective understanding about agent behavior in heterogeneous traffic is nascent, authors are often forced to leverage simulation models to develop their decision-support prototypes.[37-39] The incorporation of Bayesian reasoning within such simulations enables decision-support systems to be designed that more faithfully capture and acknowledge the associated uncertainties; however, this approach is utilized sparingly within existent literature.

Therefore, in the remainder of this section, we illustrate the development of such an approach on a lane-changing problem, arguably among the most complex maneuvers. To counter common critiques of game-theoretic reasoning and more faithfully represent the underlying uncertainties, an adversarial risk analysis (ARA) methodology[40] is codified herein. Empirical results of its behavior in a simulated environment are explored.

5.1 | ADS lane-changing via adversarial risk analysis

Consider a two-lane road wherein a level-3 to level-5 ADS (i.e., player A) is the target vehicle facing a lane-changing decision. The objective is to embed prescriptive decision making within this ADS. A Stackelberg-game structure is adopted such that our ADS is the leader, and an MV is the follower. The ADS takes an action that is observed by (the driver of) a manned-lag vehicle in the adjacent lane. This MV (i.e., player M) chooses his best response based on this observation. The interaction begins when player A receives stimuli that triggers a potential lane change and ends when both vehicles have made and executed their respective decisions.

Figure 5 represents this problem as a bi-agent influence diagram (BAID)[41] whose dynamics can be summarized as follows. The target vehicle A makes a lane-changing decision $a \in \mathcal{A}$, which is observed by the lag vehicle M, who subsequently takes some action $m \in \mathcal{M}$. Each vehicle's decision is based upon their understanding of the roadway environment's state, denoted by a latent variable $\theta \in \Theta$, which may summarize multitudinous structural, physical, and perceptual conditions (visibility, obstacle presence, etc.). The ADS infers θ via sensor data denoted by Y_A. Similarly, (the driver of) the lag vehicle uses analogous information, Y_M, to construct their beliefs about θ. Each of these inputs are non-deterministic, assuming values $y_A \in \mathcal{Y}_A$ and $y_M \in \mathcal{Y}_M$, thereby enabling the tractable modeling of perception error. Interactions between the decisions a and m, in conjunction with θ, lead to a probabilistic outcome S that assumes a value $s \in \mathcal{S}$. Players A and M receive respective utilities u_A and u_M based upon the corresponding vehicle action, as well as the realized outcome s.

To identify the ADS's optimal decision, that is, $a^*(y_A) \in \mathcal{A}$, we must codify player A's utility function and beliefs. The utility from making decision $a \in \mathcal{A}$ and inducing an outcome $s \in \mathcal{S}$ is denoted by $u_A(a, s)$. The term $p_A(\theta|y_A)$ represents player A's conditional probability density (mass) function on the roadway-environment's state given the sensor data $y_A \in \mathcal{Y}_A$, whereas player A's conditional probability density (mass) function associated with the MV decision $m \in \mathcal{M}$ conditioned on $a \in \mathcal{A}$ and $y_M \in \mathcal{Y}_M$ is denoted by $p_A(m|a, y_M)$, and the ADS beliefs about $y_M \in \mathcal{Y}_M$ given θ is denoted by $p_A(y_M|\theta)$. Finally, $p_A(s|m, a, \theta)$ represents player A's posterior probability density (mass) function for the outcome $s \in \mathcal{S}$ given the decisions m and a as well as the roadway-environment's state θ. The cumulative distribution functions associated with each of these probability density (mass) functions are denoted by $P_A(\cdot)$. Given these inputs, the ADS's objective is to find some $a \in \mathcal{A}$ that maximizes the conditional expected utility, $\mathbb{E}_A[u_A(a, s)|y_A]$. Therefore, the ADS wishes to identify the decision $a^*(y_A)$ that satisfies[**]:

$$a^*(y_A) = \arg\max_{a \in \mathcal{A}} \left[\mathbb{E}_A[u_A(a, s)|y_A]\right] = \arg\max_{a \in \mathcal{A}} \left[\int_S u_A(a, s) \mathrm{d}P_A(s|a, y_A)\right]. \quad (1)$$

[**]We denote the expected value of a function of random variables with the Riemann–Stieltjes integral to maintain generality. Our formulation assumes numeric random variables; however, relatively simple modifications can be made to accommodate categorical quantities.

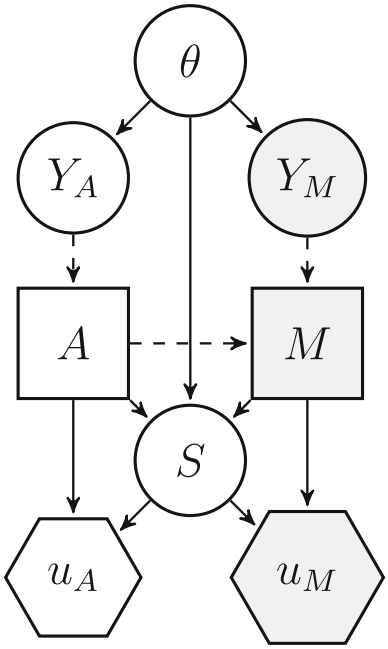

FIGURE 5 BAID representation of a two-lane game.

Since $P_A(s|a, y_A)$ is not readily available to the ADS, player A must instead maximize an equivalent representation of the expected utility

$$\mathbb{E}_A[u_A(a,s)|y_A] = \int_\Theta \int_{y_M} \int_{\mathcal{M}} \int_S u_A(a,s) \mathrm{d}P_A(s|m,a,\theta) \mathrm{d}P_A(m|a,y_M) \mathrm{d}P_A(y_M|\theta) \mathrm{d}P_A(\theta|y_A). \tag{2}$$

From a decision-analytic perspective, each of the elements in Equation (2) are standard, with the exception of $P_A(m|a, y_M)$, which has a complicating strategic component related to how the MV solves its decision problem. Therefore, additional analysis is required to form an estimate $\hat{P}_A(m|a, y_M)$ that can be utilized to resolve the ADS's decision problem. We provide such an analysis, in accordance with standard ARA procedures.[40]

More specifically, estimating $\hat{P}_A(m|a, y_M)$ requires the ADS to assess the MV's decision problem. This is similar to that of Equation (2), but is based upon information known only to the MV. To build intuition about the ARA solution procedure, suppose, for the moment, that the ADS can parameterize the MV's problem exactly. That is, the MV's utility function and probability judgements are known with certainty by the ADS. Based on these functions and, assuming the lag vehicle's driver is an expected utility maximizer, upon observing a and y_M, player M will select an action $m^*(a, y_M)$ that maximizes the conditional expected utility, $\mathbb{E}_M[u_M(m, s)|a, y_M]$, so that

$$m^*(a, y_M) = \arg\max_{m \in \mathcal{M}} \left[\mathbb{E}_M[u_M(m,s)|a,y_M]\right] = \arg\max_{m \in \mathcal{M}} \left[\int_S u_M(m,s) \mathrm{d}P_M(s|m,a,y_M)\right],$$

or, equivalently, using the standard influence diagram reductions in Shachter,[42]

$$m^*(a, y_M) = \arg\max_{m \in \mathcal{M}} \left[\int_\Theta \int_S u_M(m,s) \mathrm{d}P_M(s|m,a,\theta) \mathrm{d}P_M(\theta|y_M)\right].$$

Without loss of generality, this reduction excludes the MV's beliefs about y_A given a and y_M. Such a simplification is a modeling choice rooted in the assumption that an MV is unable to conceptualize the ADS sensor outputs (e.g., from a CNN) and is likely to ignore them[††].

[††]This assumption can be modified with relatively minor modification to the resultant formulation; only an additional distribution, with attendant uncertainty, must be added about y_A given a and y_M.

However, the ADS does not know the MV's parameterization with certainty, and this equation cannot be resolved exactly. However, if uncertainty about this parameterization is modeled in a Bayesian manner, we may codify the ADS's perception of the MV's utility and beliefs via a random utility function $U_M(m,s)$ and random probability distributions $\Pi_M(\cdot)$ that lead to

$$M^*(a,y_M) = \arg\max_{m\in\mathcal{M}} \left[\int_\Theta \int_S U_M(m,s)\mathrm{d}\Pi_M(s|m,a,\theta)\mathrm{d}\Pi_M(\theta|y_M) \right].$$

As a function of a random utility function and random probability distributions, it is clear that $M^*(a,y_M)$ is a random variable as well. Moreover, by simulating random variates from the random utility function and random probability distributions, the ADS can use these equations to form an estimate $\hat{P}_A(m|a,y_M)$ of $P_A(m|a,y_M)$. In turn, this quantity can then be utilized to resolve Equation (2).

5.2 | Empirically evaluating the lane-changing methodology

This section illustrates the foundational modeling elements set forth in the previous section and provisions a basic empirical evaluation of the ARA approach with the simple simulation environment described in Section 2.4. Additional details can be found in the work of Naveiro et al.[43]

5.2.1 | Driving scene overview

In a general setting, the ADS's actions may be multidimensional, including features like the vehicle's acceleration or the angle of the steering input. However, for illustration, we consider a unidimensional setting with $\mathcal{A} = \{a_1, a_2, a_3\}$ such that a_1, a_2, and a_3 correspond to *change lane*, *remain in lane*, and *perform an emergency maneuver*, respectively. Similarly, the MV's action space is assumed unidimensional with $\mathcal{M} = \{m_1, m_2, m_3\}$ such that m_1, m_2, and m_3 correspond to *accelerate*, *decelerate*, and *change lane*. With regard to the outcomes, we set $\mathcal{S} = \{s_1, s_2, s_3, s_4, s_5\}$ whereby the s_i-values are respectively associated with a *major accident* (i.e., the MV and ADS are essentially destroyed with fatalities occurring in both vehicles); *minor accident* (i.e., minor physical damage to the vehicles and passengers); *safely executed interaction* (i.e., no physical damage or injuries); *pedestrian casualty* (i.e., emergency diversion required with at-risk pedestrians passing away) and *crash with obstacle* (i.e. injuries to all ADS passengers and physical damage to the vehicle itself but no damage to the MV). Finally, a four-dimensional roadway environment is considered such that $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$. The element θ_1 captures the pavement's moisture condition (i.e., wet or dry), whereas θ_2, θ_3, and θ_4 respectively represent the quantities of people in the ADS, occupants in the MV, and pedestrians at risk if an emergency maneuver must take place.

5.2.2 | ADS preference and belief models

As discussed in Section 4, multitudinous objectives may be considered in ADS operations. Herein, we consider *internal safety*, *external safety*, and *trip duration* as relevant. These consequences are described via a vector $c_A(s,a) = (c_{A,1}(s,a), \ldots, c_{A,8}(s,a))$ with elements that respectively refer to the number of internal injuries, the number of internal fatalities, the proportion of ADS damage, the number of external injuries, the number of external fatalities, the proportion of MV damage, the number of pedestrians killed, and the ADS's speed. For example, when there is a major crash (i.e., s_1 occurs), θ_2 people will die in the ADS and θ_3 will perish in the MV; both vehicles will be totally damaged but no pedestrian casualties will occur. The vehicle's action (e.g., speed selection) directly affects some, but not all of these consequences. The preference model is characterized by a constant absolute risk averse (CARA) utility function that additively aggregates these values. Namely, the ADS's utility is calculated as $u_A(s,a) = 1 - \exp\left(\rho_A \sum_{i=1}^{8} w_{A,i} c_{A,i}(s,a)\right)$, wherein ρ_A is the ADS's risk aversion coefficient, and the weights $w_{A,i}$ homogenize the criteria. This function is used, in conjunction with the ADS's subjective beliefs, to determine her expected utility. To do so, the ADS must formalize three non-strategic probabilistic quantities, that is, beliefs about θ given y_A; beliefs about s given m, a, and θ; and beliefs about y_M given θ. In our setting, the pavement's moisture condition (i.e., $\theta_1 = 0$ for dry, 1 for wet) is predicted via the sensor y_A, which is assumed

to be a machine-learning classifier (e.g., a CNN) characterized by 95% accuracy[‡‡], that is, $p(\theta_1 = 0|y_A = 0) = 0.95$. The remaining elements in θ are assumed to be known. A tabular conditional probability distribution defines $p_A(s|m, a, \theta)$ having, for example, probabilities equal to $\frac{2}{3}$, $\frac{1}{6}$, $\frac{1}{6}$ 0 and 0 for s_1 through s_5, respectively, when $m = m_1$, $a = a_1$, and $\theta_1 = 0$. Finally, for illustration, we assume that the ADS believes that the MV observation y_M is equal to y_A.

5.2.3 | MV random preference and belief models

As detailed herein, the ARA requires the ADS to formalize their beliefs about the MV's decision problem to estimate $p_A(m|a)$. The ADS assumes the MV bases their decision on the internal safety and trip duration objectives; thus, the driver of an MV is supposed to act less altruistically than an ADS. A CARA preference model is similarly assumed for the MV. Namely, we have $u_M(s, m) = 1 - \exp\left(\rho_M \sum_{i=1}^{4} w_{M,i} c_{M,i}(s, m)\right)$, wherein ρ_M and $w_{M,i}$ again represent respectively the risk-aversion coefficient and homogenizing weights. However, for the MV's problem, the ADS is uncertain about these parameters. The ADS's beliefs about $(w_{M,1}, w_{M,2}, w_{M,3}, w_{M,4})$ are characterized by a Dirichlet distribution, whereas ρ_M is characterized by a uniform distribution. The MV's utility is therefore a random variable depending on these quantities. Furthermore, to estimate the MV's expected utility, the ADS must also formalize beliefs about the probability models discussed in Section 5.1. Notably, $\Pi_M(\theta|y_M)$ is a complex quantity to determine directly because the ADS must estimate the cognitive processing of the MV's driver based on the impulse he receives. One promising approach is to base this quantity on $p_A(\theta|y_A)$ with some additional attendant uncertainty. For example, we root $\Pi_M(\theta_1|y_M)$ on $p(\theta_1|y_A)$ such that its values are reflected as means of a beta distribution. The ADS similarly characterizes $\Pi_M(s|m, a, \theta)$ based on the $p_A(s|m, a, \theta)$-values, but with Dirichlet distributions. For example, when we condition on m_1, a_1 and $\theta_1 = 0$, we use a Dirichlet distribution having parameters $100 \cdot \left(\frac{2}{3}, \frac{1}{6}, \frac{1}{6}, 0, 0\right)$ such that its mean is $\left(\frac{2}{3}, \frac{1}{6}, \frac{1}{6}, 0, 0\right)$. Collectively, these beliefs describe a hierarchical model over the MV's expected utility whose maximum over $m \in \mathcal{M}$ is used to estimate $\hat{p}_A(m|a)$.

5.2.4 | Testing configuration

Experimentation is set forth utilizing the aforementioned simulation model (coded in Python) with modifications to estimate the probabilities of different MV reactions given the ADS decision and, based on these probabilities, a routine maximizing the ADS's expected utility. The ADS utility weights are set to $(0.03, 0.21, 0, 0.03, 0.21, 0, 0.21, 0.31)$, and ρ_A is set to 0.5. The number of passengers in the ADS ranges from 0 to 5, whereas the number of passengers in the MV ranges from 1 to 5. The MV utility weights follow a Dirichlet distribution with parameters $\alpha(0.1, 0.5, 0.05, 0.35)$, and ρ_M follows a uniform distribution over (0.5, 1.5). Finally, the number of people in the street ranges from 0 to 6. Observe that, from a global perspective, this is a quite risky driving scene, in the sense that both the ADS and the MV highly value speed over safety. Hyperparameter α is set to 100.

5.2.5 | Results

An experiment was conducted to verify the soundness of the approach such that (1) a single pedestrian was present and, (2) both the ADS and MV had a single occupant. Monte Carlo simulation was performed to estimate $\hat{p}_A(m|a)$ as presented in Table 2. Notice that the ADS beliefs about the MV action are conditioned on their own action. Recall that the MV actions m_1, m_2, and m_3 refer respectively to accelerate, decelerate and and change lanes, while those of the ADS a_1, a_2, and a_3 refer to change lanes, stay in lane, and emergency stop.

The derived results confirm intuition. Notably, should the ADS change lanes (i.e., take action a_1), she believes the MV will most likely decelerate or (with lesser probability) change lanes as well. In turn, when the ADS decides to remain in the same lane (i.e., a_2), it is sure that the MV will accelerate. Finally, when the ADS decides to make an emergency stop (i.e., a_3), the ADS is also sure that the MV will accelerate. Utilizing this model, the ADS can explore how the MV's behavior

[‡‡]We recognize that the level of accuracy demonstrated is not sufficient for real-world applications. Nevertheless, it serves its purpose as an illustrative example.

TABLE 2 ADS estimate of $p_A(m|a)$.

		MV actions		
		m_1	m_2	m_3
ADS action	a_1	0.000	0.860	0.140
	a_2	1.000	0.000	0.000
	a_3	1.000	0.000	0.000

TABLE 3 Conditional proportions of ADS's preferred action and resulting outcomes based on true latent state.

	ADS actions			Interaction outcomes				
θ_1	a_1	a_2	a_3	s_1	s_2	s_3	s_4	s_5
0	0.95	0.05	0	0.158	0.158	0.681	0.0	0.003
1	0.05	0.95	0	0.017	0.017	0.682	0.0	0.285

affects her expected-utility-maximizing decision. A collection of simulations of the interaction was performed under each value of θ_1 (unobserved by the drivers). Table 3 presents the proportion of simulations in which each ADS action was the expected-utility-maximizer. Likewise, it also presents the proportion of outcomes, provided the ADS selects the attendant expected-utility-maximizing action. Therein, it can be observed that, when the pavement is dry, the ADS generally prefers to change lanes, whereas, when the pavement is wet, she tends to act conservatively by staying in place. The resulting disparity in actions is rooted in the error of sensor y_A, highlighting the need for effective measurement tools. Moreover, the proportion of outcomes suggests that, despite the hazardous environment, more often than not the interaction occurs safely.

5.3 | Open problems in heterogeneous-traffic operations

Our initial results with simple simulators have been promising and have prompted additional experimentation higher up the ADS testing ladder,[44] but numerous and far-ranging problems must be resolved before this methodology may be utilized in real-world ADSs. Although the framework provided herein is incredibly flexible, as a probabilistic graphical model it suffers from standard limitations. Our empirical exploration examined a BAID characterized by a collection of conditional probability tables and, whereas discretization is commonly used in such problems to facilitate computations, it inherently limits the fidelity of the model. In principle, the distributions in Figure 5 may be arbitrary but, in so doing, inference and optimization along the BAID become increasingly difficult. Describing uncertainties with conditional Gaussians is promising and computationally feasible;[45] however, since such an approximation may still not be faithful enough to reality, alternative characterization of increasing complexity may be necessary.[46] Similarly, the agents' decision variables may be continuous and, whereas discretization enables rapid computation, it only provides an optimal answer to an approximation of the true problem. Such an approach is common practice within decision analysis, but its efficacy in the ADS setting is an open problem. Ultimately, additional, real-world experimentation is the only way to better understand such tradeoffs.

Besides, algorithmic efficacy in the decision layer must be balanced with the computing needs of the perception and forecasting layers. Computational resources onboard an ADS are limited, and computing time must be balanced across algorithms to ensure proper ADS performance. A computer vision algorithm used for y_A can be designed independent of the algorithms resolving Figure 5 but, in reality, the methods work in tandem and must be developed accordingly. For example, within our experimentation, the simulations informing $\hat{p}(m|a)$ constituted roughly 95% of the total computation time, implying that the identification of more efficient estimation methods may be requisite for computationally constrained environments. Moreover, since more complex conditions (e.g., 1 ADS + n MVs) demand even more computational effort, for ARA to remain feasible in such problems, the development of novel approximation schemes may be required.

6 | PROTECTING ADSs FROM ADVERSARIAL DATA

As discussed in Section 2.3, and sketched in Sections 3 and 5, sensor input is fundamental to ADS operations. This raw information is utilized by all manner of perception algorithms for state estimation. Such estimates are subsequently used as input for downstream decision making algorithms. Given the degree to which sensor data and perception algorithms affect ADS operations, it is essential that these components are robust and reliable, especially in light of contemporary, adversarial machine learning (AML) threats.

Although state-of-the-art machine-learning (ML) algorithms perform extraordinarily well on trusted data, recent research has proven them vulnerable to data manipulation, for example, the corruption of input training data.[47] Adversarial examples, data instances strategically designed to thwart an algorithm, typify the efficacy of data manipulation. These attacks have proven devastating in numerous computer-vision tasks.[48] The simplest adversarial examples modify an image imperceptibly to the human eye, but induce classification errors in even the most well-tested algorithms.[49] Such vulnerabilities have particular salience in an ADS setting given the widespread use of computer vision and the potential negative repercussions of misclassifications. Multiple authors have demonstrated how attacks may be tailor-made to thwart proper ADS perception (for a review, see Huq et al.[3]). Notably, Zhang et al.[50] proposed DeepRoad, a GAN-based approach to generate test images for real world driving scenes, and Zhou et al.[51] introduced DeepBillboard to generate real-world adversarial billboards that can trigger ADS steering errors. Additional AML attacks tailored to the ADS setting have been developed by Cao et al.,[52] Boloor et al.,[53] and Deng et al.[54] The effects of such attacks have proven highly disruptive to ADS performance.

Therefore, to enhance the security of an ADS's machine-learning pipeline, foundational and domain-agnostic AML research is required. As noted by Fan et al.,[55] a framework that principally guarantees robust ML against adversarial manipulations is lacking; initial research addressing this problem is promising, but it is not a panacea. Game theory[56] is the, sometimes implicit, prevailing paradigm used to address the confrontation between adversaries and learning-based systems. However, this approach often requires a strong variant of the common knowledge assumption which, from a conceptual standpoint, is of dubious veracity in security applications.[57] An alternative approach to model robustification, that is, *adversarial training* (AT), is based on solving a bilevel optimization problem that sets the objective function equal to the empirical risk under worst-case data perturbations[§§].[58] Unfortunately, as with other robust optimization techniques, it is susceptible to over-conservatism.

Since these (and other) existent methods have their relative advantages and disadvantages, recent research has urged traditional, norm-based evaluation be set aside in favor of more realistic attack models.[59] Additionally, because opponent models are implicit within the aforementioned approaches, one cannot deviate far from their assumptions without invalidating the underlying methodology. Therefore, in an effort to address such rigidity and increase opponent-model realism, we discuss a Bayesian approach to this challenge. In the remainder of this section, we summarize and illustrate this approach for a general classification task, and explore requisite ADS-specific extensions.

6.1 | A Bayesian approach to threatened classification

Since an ADS manufacturer controls experimentation, it is better able to protect against data manipulation during system development than after its deployment. Namely, although it is reasonable to assume that the training data used to develop an ADS classifier is trustworthy, the operational data is likely threatened. Therefore, because the data-generation processes are distinct in the training and operational environments, it is simply ill-advised to train a classifier without taking these dynamics into account.

The Bayesian adversarial learning (BAL) paradigm set forth in Ye and Zhu[60] provides a general framework for addressing these disparate data-generation processes. Assuming clean training data is available, the methodology produces artificially threatened data that mimics how an attacker would perturb it. This allows the Defender (i.e., the ADS perception system) to essentially map the training environment to the operational environment. The artificially threatened data is subsequently leveraged to construct a classifier. Assuming the process mapping clean data to threatened data faithfully embodies the true dynamics, the resulting classifier should be of higher quality than another trained on clean data alone.

The BAL framework considers a Defender using a model parameterized by θ who has clean training data, $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$, wherein x_i is a vector of features and y_i is a label. The Defender knows that, under operational conditions,

[§§] Note the implicit game-theoretic flavor.

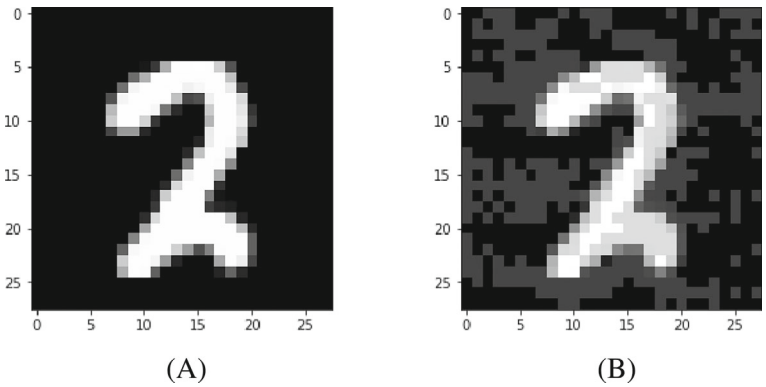

FIGURE 6 Example effect of FGSM attacks on the MNIST data. (A) Original image. (B) Perturbed image.

an attacker would modify $\mathcal{D}$ to $\tilde{\mathcal{D}}$; however, they are not certain exactly how these perturbations would occur. Therefore, rather than canonically computing a posterior over θ, the authors suggest computing a *robust adversarial posterior distribution* via

$$p(\theta|\mathcal{D}) = \int p(\theta, \tilde{\mathcal{D}}|\mathcal{D})\,\mathrm{d}\tilde{\mathcal{D}} = \int p(\theta|\tilde{\mathcal{D}})p(\tilde{\mathcal{D}}|\mathcal{D})\,\mathrm{d}\tilde{\mathcal{D}}.$$

This robust adversarial posterior can then be used for class prediction using standard calculations. Namely, given some x_j, beliefs about y_j are captured by

$$p(y_j|x_j, \mathcal{D}) = \int p(y_j|x_j, \theta)p(\theta|\mathcal{D})\,\mathrm{d}\theta.$$

If samples from the robust adversarial posterior, that is, $p(\theta|\mathcal{D})$, are available, this integral can be calculated using Monte Carlo integration. Notably, samples from $p(\theta|\mathcal{D})$ can be obtained through Gibbs sampling, as Ye and Zhu[60] explain. If paired with a utility function, similar calculations may be leveraged to maximize the learner's posterior predictive utility.

Although the BAL framework is well-defined, the attacker model (i.e., $p(\tilde{\mathcal{D}}|\mathcal{D})$) is generically specified. A continuum of models can therefore be constructed by varying assumptions about this quantity. Rios Insua et al.[61] and Rios Insua et al.[62] show that the framework subsumes AT when $p(\tilde{\mathcal{D}}|\mathcal{D})$ is a degenerate distribution. Ye and Zhu[60] provide another simple model wherein (1) the attacker may only perturb features (i.e., x_i) but not labels (i.e., y_i), and (2) the probability of an attack is a function of the attacker's risk-reward balance. Alternatively, Rios Insua et al.[61] and Rios Insua et al.[62] more fully explore the flexibility of the BAL framework by leveraging ARA to develop $p(\tilde{\mathcal{D}}|\mathcal{D})$. This pairing further generalizes the BAL framework by allowing a wide array of (even deterministic) attacks to be formally encoded as the attacker's model.

6.2 | Empirically evaluating the robustified classifier

To illustrate the efficacy of the ARA approach to BAL, we discuss its application to a well-known, computer-vision benchmark, that is, the MNIST dataset.[63] The defender aims to correctly identify digits (i.e., from 0 to 9), but an attacker perturbs the MNIST data to thwart this classification. A direct analog of this example in the ADS settings relates to the confounding of a yield sign with a stop sign. For illustration, we assume the attacker perturbs images using the fast gradient sign method (FGSM) or the projected gradient descent approach (PGD) developed in Goodfellow et al.[48] and Madry et al.,[58] respectively. Figure 6 highlights the efficacy of constructing adversarial examples in this manner. Whereas both the original and perturbed image appear to depict the same digit, using a benchmark CNN, the former is correctly classified as a 2 but the latter is incorrectly identified as a 7.

Four AML defense methods are used to robustify our baseline CNN against these two attack methods. Namely, adversarial training, adversarial logit pairing (i.e., ALP[64]) and undefended (i.e., None) approaches are compared against the

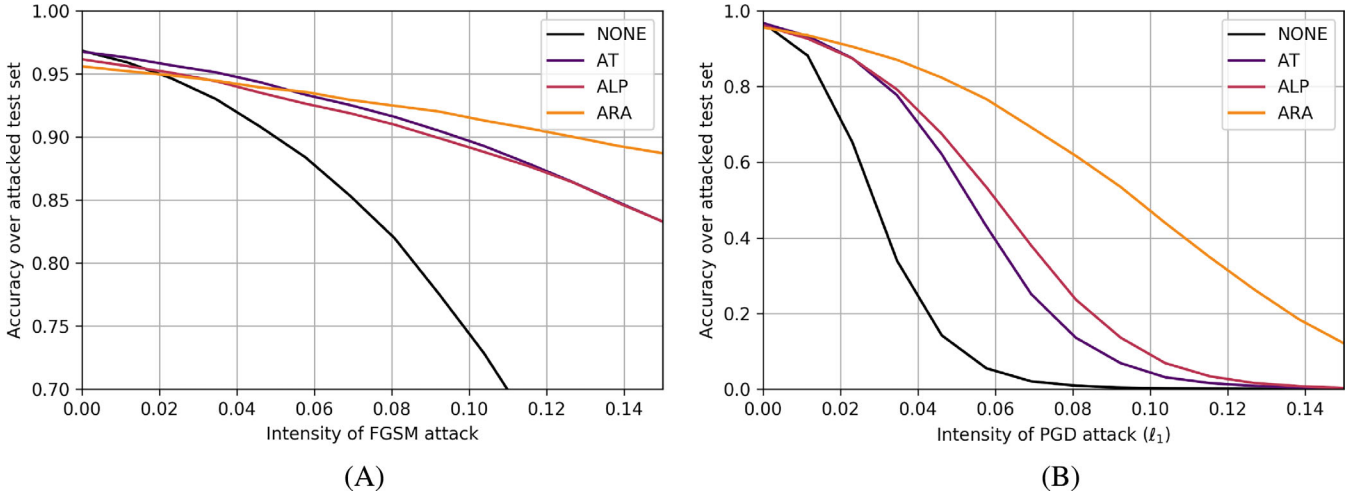

FIGURE 7 Classifier accuracy under three defence mechanisms and two attacks. (A) FGSM attack. (B) PGD Attack under l_1-norm.

previously described ARA methodology. Figure 7 plots the *security evaluation curves* for each defense, which depict the accuracy of the defender model under different attack intensities[¶¶].

Inspection of Figure 7 immediately reveals noteworthy behavioral patterns across the four examined defenses. At low attack intensities, all four defenses perform comparably, but the AML defenses appear to yield marginally lower accuracies. Alternatively, as attack intensity increases, the performance of the undefended technique rapidly deteriorates, thereby illustrating the essential nature of AML defenses. The undefended model's accuracy on the untainted data is 98% and quickly degrades to 75% under FGSM attack at an intensity of 0.1. The three robustified approaches mitigate this degradation, but with varying degrees of success. The AT and ALP defenses perform comparably under FGSM attack (Figure 7A), but the AT defense degrades quicker under the PGD attack (Figure 7B). The ARA approach generally appears more robust than the AT and ALP defenses at higher intensities for both attacks. For example, a 0.1 FGSM-attack intensity induces AT and ALP accuracies of 90% but an ARA accuracy of 92%; this difference increases further with FGSM-attack intensity. The disparities are even more pronounced under PGD attack. Although such attacks degrade all defenses, higher attack intensities are required to affect the ARA defense than the AT and ALP defenses.

6.3 | Open adversarial machine learning problems

The previous section illustrates the efficacy of the ARA approach in defending an ML algorithm. However, the developed methodology is not without its shortcomings. Notably, the aforementioned framework essentially simulates the attacker problem to forecast attacks, and utilizes this information to optimize the defender's decision. This entails a non-trivial amount of computational resources and effort; two quantities in limited supply during ADS operations.

Therefore, future ARA research of an algorithmic nature is a crucial need. A single-stage approach, augmented probability simulation[65] is promising in that it combines the Monte Carlo sampling and optimization routines. However, should this approach prove computationally infeasible as well, alternative means may be required. Such alternatives may include expedient heuristics for the attacker's problem or approximation techniques that regress the attacker's best response function in a metamodeling sense.

Future empirical research is required to determine which of the aforementioned algorithmic directions is most promising. Intuition implies that, at some point, standard ARA practices will need to be amended to accommodate more sophisticated AML problems but, without additional empirical testing, this threshold cannot be determined. Moreover, since differing ML algorithms require varying degrees of computational effort, this threshold is likely to be context-dependent, implying that each ML algorithm may require its own bespoke analysis.

[¶¶] Larger attack intensities imply more powerful attacks; for more information see Rios Insua et al.[62]

7 | CONCLUSION

ADS technology is rapidly advancing, but it will plateau if outstanding technological, methodological, regulatory and ethical problems remain unaddressed.[4] The problems described herein are just a subset of those for which statisticians are well-equipped to address. As such, we described initial research aimed at dealing with these issues, and also detailed a few open challenges for each.

Nevertheless, the breadth of statistical problems in an ADS setting extends well beyond those enumerated herein. For example, it is well-known that ADS technology depends upon deep learning;[13] such neural network models were mentioned numerous times within this manuscript. Although traditional deep learning approaches have enabled a boom in ADS development, McAllister et al.[10] detail the benefits of Bayesian deep learning for ADS applications. Rather than outputting precise point predictions, a Bayesian approach allows uncertainty to propagate throughout the ML pipeline to better inform downstream decisions. However, most deep-learning training algorithms are based upon the maximum-likelihood-estimation tradition. Efficient Bayesian integration methods for deep neural networks have yet to be identified.[66] Therefore, this general statistical challenge is particularly relevant to ADS applications.

Furthermore, from an application perspective, additional ADS research issues are of relevance to statisticians as well (e.g., see Figure 1). Automobile insurance is an exemplar. The massive adoption of ADS technology will revolutionize the industry by modifying the operational environment and provisioning additional data streams. Notably, architectures akin to that presented in Figure 2 allow for the development of continuous risk monitoring in support of service-level agreements that promote safer ADS operations, for example, by incentivizing through cheaper insurance policies preserving a risk indicator below a certain value. The resolution of such problems will invariably require intradisciplinary collaboration between several statistical subfields.

Finally, statisticians must also support other empirical ADS challenges. Properly designed, executed, and analyzed experiments are required for developing ADS software; understanding driver behavior in heterogeneous traffic; modernizing future driver education programs; ensuring the interpretability of machine learning algorithms used in automated driving systems; improving rare-events modeling; addressing the challenges of data privacy and security in the collection, storage, and processing of large amounts of sensitive driving data; and forging sound ADS regulation; among other problems. The need for sound statistical analysis is paramount. Without it, ADS technology cannot reach its full potential.

ACKNOWLEDGMENTS

The authors acknowledge the support of the National Science Foundation under Grant DMS-1638521 to the Statistical and Applied Mathematical Science Institute (SAMSI), NC, USA. David Rios Insua is supported by the AXA-ICMAT Chair. This work supported by the EU's Horizon 2020 projects 815003 Trustonomy and 101021797 STARLIGHT and Grants from the AMALFI FBBVA project, the Spanish Ministry of Research PID2021-124662OB-I00, the Air Force Scientific Office of Research (AFOSR) award FA-9550-21-1-0239, and the AFOSR European Office of Aerospace Research and Development award FA8655-21-1-7042. William N. Caballero is partially supported by the AFOSR Grant 21RT0867. We are grateful for the comments of two anonymous reviewers.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Roi Naveiro https://orcid.org/0000-0001-9032-2465

REFERENCES

1. Burns L, Shulgan C. *Autonomy: The Quest to Build the Driverless Car—And How It Will Reshape Our World*. ECCO; 2019.
2. AMSTAT. Statistical science improving transportation; 2012. https://www.amstat.org/asa/files/pdfs/StatSig/StatSigTransportation.pdf
3. Huq N, Vosseler R, Swimmer M. Cyberattacks against intelligent transportation systems. Technical report. Trend Micro; 2017.
4. Jeffrey HH, amd E, Vijay B, Kendall A. Reimagining an autonomous vehicle. arXiv preprint arXiv:2018.05805, 2021.
5. BMW. The path to autonomous driving; 2023. https://www.bmw.com/en/automotive-life/autonomous-driving.html
6. Society of Automotive Engineers. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Technical report. SAE; 2018. doi:10.4271/j3016&uscore;201806
7. Consumer Reports. Guide to adaptive cruise control; 2019a. https://www.consumerreports.org/car-safety/adaptive-cruise-control-guide/
8. Consumer Reports. Guide to lane departure warning & lane keeping assist; 2019b. https://www.consumerreports.org/car-safety/lane-departure-warning-lane-keeping-assist-guide/

9. Caballero WN, R D, Banks D. Decision support issues in automated driving systems. *Int Trans Oper Res*. 2021;30:1216-1244.
10. McAllister R, Gal Y, Kendall A, et al. Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence; 2017:4745–4753.
11. Wang Q, Ayalew B, Weiskircher T. Predictive maneuver planning for an autonomous vehicle in public highway traffic. *IEEE Trans Intell Transp Syst*. 2018;20(4):1303-1315.
12. Bojarski M, Choromanska A, Choromanski K, et al. VisualBackProp: visualizing CNNs for autonomous driving. arXiv preprint arXiv:1611.054182, 2016.
13. Grigorescu S, Trasnea B, Cocias T, Macesanu G. A survey of deep learning techniques for autonomous driving. *J Field Robot*. 2020;37(3):362-386.
14. NTSB. Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator. Technical report. National Transportation Safety Board; 2020.
15. Czarnecki K. Operational design domain for automated driving systems taxonomy of basic terms. Technical report. University of Waterloo; 2018.
16. Insua DR, Caballero WN, Naveiro R. Managing driving modes in automated driving systems. *Transp Sci*. 2022;56:1259-1278.
17. West M, Harrison J. *Bayesian Forecasting and Dynamic Models*. Springer; 2006.
18. Claussmann L, Revilloud M, Gruyer D, Glaser S. A review of motion planning for highway autonomous driving. *IEEE Trans Intell Transp Syst*. 2019;21:1826-1848. doi:10.1109/tits.2019.2913998
19. Bianchi S. Trustonomy: building the acceptance of automated mobility. Intelligent Transport; 2018. https://www.intelligenttransport.com/transport-articles/99585/trustonomy-building-the-acceptance-of-automated-mobility/
20. French S, Insua DR. *Statistical Decision Theory*. Edward Arnold; 2000.
21. Awad E, Dsouza S, Kim R, et al. The moral machine experiment. *Nature*. 2018;563(7729):59-64.
22. Lin P. Why ethics matters for autonomous cars. In: Maurer M, Gerdes J, Lenz B, Winner H, eds. *Autonomous Driving*. Springer; 2016:69-85.
23. Keeney RL. Ethics, decision analysis, and public risk. *Risk Anal*. 1984;4(2):117-129.
24. Caballero WN, Naveiro R, R D. Modeling ethical and operational preferences in automated driving systems. *Decis Anal*. 2022;19(1):21-43.
25. Keeney RL, Raiffa H, Meyer RF. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press; 1993.
26. Keeney RL, Gregory RS. Selecting attributes to measure the achievement of objectives. *Oper Res*. 2005;53(1):1-11.
27. Wakker P. *Prospect Theory for Risk and Ambiguity*. Cambridge University Press; 2008.
28. Kim R, Kleiman-Weiner M, Abeliuk A, et al. A computational model of commonsense moral decision making. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society; 2018:197–203.
29. Lee D, Hess DJ. Regulations for on-road testing of connected and automated vehicles: assessing the potential for global safety harmonization. *Transp Res Part A Policy Pract*. 2020;136:85-98.
30. Stange V, Kühn M, Vollrath M. Safety at first sight?—Manual drivers' experience and driving behavior at first contact with level 3 vehicles in mixed traffic on the highway. *Transp Res Part F Traffic Psychol Behav*. 2022;87:327-346.
31. Hult R, Zanon M, Gros S, Wymeersch H, Falcone P. Optimisation-based coordination of connected, automated vehicles at intersections. *Veh Syst Dyn*. 2020;58(5):726-747.
32. Zheng Y, Wang J, Li K. Smoothing traffic flow via control of autonomous vehicles. *IEEE Internet Things J*. 2020;7(5):3882-3896.
33. Zhong Z. *Assessing the Effectiveness of Managed Lane Strategies for the Rapid Deployment of Cooperative Adaptive Cruise Control Technology*. Ph.D. thesis. New Jersey Institute of Technology; 2018.
34. Chen B, Sun D, Zhou J, Wong W, Ding Z. A future intelligent traffic system with mixed autonomous vehicles and human-driven vehicles. *Inform Sci*. 2020;529:59-72.
35. Di X, Shi R. A survey on autonomous vehicle control in the era of mixed-autonomy: from physics-based to AI-guided driving policy learning. *Transp Res Part C Emerg Technol*. 2021;125:103008.
36. Schwarting W, Pierson A, Alonso-Mora J, Karaman S, Rus D. Social behavior for autonomous vehicles. *Proc Natl Acad Sci*. 2019;116(50):24972-24978.
37. Talebpour A, Mahmassani HS. Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transp Res Part C Emerg Technol*. 2016;71:143-163.
38. Yang D, Farah H, Schoenmakers MJ, Alkim T. Human drivers behavioural adaptation when driving next to a platoon of automated vehicles on a dedicated lane and implications on ttraffic flow: a driving simulator and microscopic simulation study in the Netherlands. Proceedings of the 98th Annual Meeting of the Transportation Research Board; 2019:19-00582.
39. Li T, Guo F, Krishnan R, Sivakumar A, Polak J. Right-of-way reallocation for mixed flow of autonomous vehicles and human driven vehicles. *Transp Res Part C Emerg Technol*. 2020;115:102630.
40. Banks D, Rios J, Rios Insua D. *Adversarial Risk Analysis*. Francis Taylor; 2015.
41. Koller D, Milch B. Multi-agent influence diagrams for representing and solving games. *Games Econ Behav*. 2003;45(1):181-221.
42. Shachter RD. Evaluating influence diagrams. *Oper Res*. 1986;34(6):871-882. doi:10.1287/opre.34.6.871
43. Naveiro R, Caballero W, Insua DR. An adversarial risk analysis for heterogeneous traffic management. Technical report; 2022.
44. Xu W, Sainct R, Gruyer D, Orfila O. A decision support framework for autonomous driving in normal and emergency situations. Proceedings of the 2021 AEIT International Conference on Electrical and Electronic Technologies for Automotive (AEIT Automotive); 2021:1–6.
45. Shachter RD, Kenley CR. Gaussian influence diagrams. *Manag Sci*. 1989;35(5):527-550. doi:10.1287/mnsc.35.5.527

46. Song L, Fukumizu K, Gretton A. Kernel embeddings of conditional distributions: a unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Process Mag*. 2013;30(4):98-111.
47. Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recognit*. 2018;84:317-331.
48. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. International Conference on Learning Representations, 2015. https://arxiv.org/abs/1412.6572
49. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. Proceedings of the International Conference on Learning Representations; 2014.
50. Zhang M, Zhang Y, Zhang L, Liu C, Khurshid S. DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems. Proceedings of the 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE); 2018:132–142; IEEE.
51. Zhou H, Li W, Kong Z, et al. DeepBillboard: systematic physical-world testing of autonomous driving systems. Proceedings of the 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE); 2020:347–358; IEEE.
52. Cao Y, Xiao C, Cyr B, et al. Adversarial sensor attack on lidar-based perception in autonomous driving. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security; 2019:2267–2281.
53. Boloor A, He X, Gill C, Vorobeychik Y, Zhang X. Simple physical adversarial examples against end-to-end autonomous driving models. Proceedings of the 2019 IEEE International Conference on Embedded Software and Systems; 2019:1–7; IEEE.
54. Deng Y, Zheng X, Zhang T, Chen C, Lou G, Kim M. An analysis of adversarial attacks and defenses on autonomous driving models. Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom); 2020:1–10; IEEE.
55. Fan J, Ma C, Zhong Y. A selective overview of deep learning. *Stat Sci*. 2021;36(2):264-290.
56. Menache I, Ozdaglar A. *Network Games: Theory, Models, and Dynamics*. Synthesis Lectures on Communication Networks. Vol 4. Springer; 2011:1-159.
57. Hargreaves-Heap S, Varoufakis Y. *Game Theory: A Critical Introduction*. Routledge; 2004.
58. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. Proceedings of the International Conference on Learning Representations; 2018. https://openreview.net/forum?id=rJzIBfZAb
59. Carlini N, Athalye A, Papernot N, et al. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705, 2019.
60. Ye N, Zhu Z. Bayesian adversarial learning. Proceedings of the 32nd Conference on Neural Information Processing Systems; 2018:6892–6901; Curran Associates Inc.
61. Rios Insua D, Naveiro R, Gallego V, Poulos J. Adversarial machine learning: Bayesian perspectives. *J Am Stat Assoc*. 2023. doi:10.1080/01621459.2023.2183129
62. Rios Insua D, Naveiro R, Gallego V. Perspectives on adversarial classification. *Mathematics*. 2020;8(11):1957.
63. LeCun Y. The MNIST database of handwritten digits; 1998. http://yann.lecun.com/exdb/mnist/
64. Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing. arXiv preprint arXiv:1803.06373, 2018.
65. Ekin T, Naveiro R, Ríos Insua D, Torres-Barrán A. Augmented probability simulation methods for sequential games. *Eur J Oper Res*. 2022;306:418-430. doi:10.1016/j.ejor.2022.06.042
66. Gallego V, Rios Insua D. Current advances in neural networks. *Annu Rev Stat Appl*. 2022;9:197-222.

How to cite this article: Caballero WN, Rios Insua D, Naveiro R. Some statistical challenges in automated driving systems. *Appl Stochastic Models Bus Ind*. 2023;39(5):629-652. doi: 10.1002/asmb.2765