Modelling Ethical Algorithms in Autonomous Vehicles Using Crash Data

Pamela Robinson,*† Landy Sun,‡ Heidi Furey,§ Ryan Jenkins,¶ Christopher R. Phillips,* Thomas M. Powers,‖ Ryan Ritterson,‡ Yuanchang Xie,** Rocco Casagrande,‡ and Nicholas G. Evans*

***Abstract*—Autonomous vehicles (AVs) could soon drive as well or better than humans. Because human drivers regularly make important ethical decisions, AVs also must be designed to behave ethically. Ethical decisions for AVs are often presented as "trolley problems," where an AV must choose between two costly outcomes. A popular method for finding ethical algorithms for AVs, the *Trolley-Preferences Method*, elicits human judgments as proxies for what an AV ought to do in emergent scenarios such as unavoidable crashes. This method, however, is limited in its understanding of a) what data are available to AVs; b) what an AV can do in a scenario, and thus what it ought to do; and c) the limits of using human preferences as a proxy for ethical action.**

In this paper we provide a proof of principle of a new method, the *Data-Theories Method*, in which naturalistic data, including vehicle crash data, is combined with philosophical ethical theory to provide a guide to action for AV algorithm design. We use this method to model three scenarios in which an AV is exposed to risk on the road, to determine possible actions for the AV. We then examine how different philosophical perspectives on agent partiality, or the degree to which an individual can act in their own self-interest, might address each scenario. Our results demonstrate that this method shows why modelling the ethics of AVs using data is essential. First, AVs may sometimes have options that human drivers do not, and designing AVs to mimic the most ethical human driver would not ensure that they do the right thing. Second, while ethical theories can often disagree about what should be done, disagreement can be reduced and compromises found with a more complete understanding of the AV's choices and their consequences. Finally, framing problems around thought experiments may elicit preferences that are divergent with what individuals might prefer once they are provided with real risks for a scenario. Our method provides a principled and empirical approach to productively address these problems and provide guidance on AV algorithm design.

***Index Terms*—autonomous vehicles, ethics, crash data, safety**

I. Introduction

AUTONOMOUS vehicles are promised to be safer, more efficient, and more cost-effective than human-driven vehicles [1]. The actions of AVs, like those of human drivers, involve ethical decisions [2]. Some of these decisions will be infrequent and momentous. E.g. the decision of who to target in an unavoidable crash [3], [4] . Others will be more commonplace and may not even seem to be ethical decisions. E.g. the decision of how closely to pass pedestrians or other vehicles [4]. Arguably, AVs must navigate both kinds of situations at least as well as humans if they are to share our roads. To reach this goal, they must be governed by navigation algorithms that produce ethically appropriate behavior.

Previous empirical work on the ethics of AVs has primarily relied on survey methods , and can be arranged into three primary themes. First, survey methods have revealed preferences individuals have for how AVs ought to behave under conditions of risk (i.e. when the probability of an outcome is known) and uncertainty (i.e. when the probability of an outcome is unknown) [5]–[7]. The second primary type of survey concerns the degree to which consumers believe AVs can be safe and useful [8], [9] and to what extent these properties should exceed those of human drivers [8], [10]. Third, perspective studies have identified the degree to which moral beliefs about AVs depend on how individuals regard themselves as situated, e.g. as pedestrian or passenger [10]–[13].

Recent attempts at using empirical methods to determine the content of ethical algorithms for AVs have elicited the public's preferences about who an AV should strike in a fatal, unavoidable, crash [5], [12], [13]. This approach, the Trolley-Preferences Method, so-called after the famous "trolley problem" thought experiment first proposed by Filippa Foot [14] , takes the form:

1) Scenario construction: Select a scenario for AVs, using ethical 'trolley cases' as inspiration. Such scenarios usually involve a choice of two actions that any driver could make, each of which have certain outcomes. For example: *hit car A and two people die, or swerve to hit car B and one person dies.*
2) Preference elicitation: Collect information about the choices people would make themselves or would prefer an AV to make.
3) Algorithm generation: Directly apply findings to generate an ethical algorithm for AVs. AVs following this algorithm will act as most people would act, as most people would prefer *most people* to act, or as most people would prefer *AVs* to act.

This approach is simple and relatively easy to use, and has the obvious advantage of producing an algorithm that conforms to most people's preferences, but is limited in a number of ways. First, it is unlikely that questions about how AVs should behave can be answered merely by polling the public. This method assumes, falsely, the AV will have access to the same information as a human driver, such as

*Department of Philosophy, University of Massachusetts Lowell, MA; †Research School of Social Sciences, Australian National University, Canberra, Australia;‡Gryphon Scientific, Takoma Park, MD;§Manhattan College, The Bronx, NY; ¶Philosophy Department, California Polytechnic State University, San Luis Obispo, CA ‖Department of Philosophy, University of Delaware, Newark, DE;**Department of Civil Engineering, University of Massachusetts Lowell, MA

Email for correspondence: nicholas_evans@uml.edu

information about the number and characteristics (such as race or occupation) of its occupants. At least in the near term, this is implausible. Moreover, restrictions on AV information may be imposed: in Germany, for example, the Federal Ministry of Transport and Digital Infrastructure prohibits the use of identifying information (gender, race, economic status, etc.) by AVs [15] . Unavoidably fatal crash scenarios are a relatively rare kind of case in which AVs put others at risk [16], [17] , whereas most accidents involve a risk of injury to the occupants, but not an assured fatality.

While AVs may lack certain information, they may also have faster reaction times, access to better information on speed, distance and accelerations of all vehicles, and more computational power to leverage all these advantages to maneuver. Given individuals frame their perceptions of AV behavior in terms of the behavior human driven vehicles [18], they may miss important ways that AVs can improve human safety and traffic efficiency. Though ethical algorithms for AVs grounded in individual preferences are not limited to assuming that AVs can only do what humans can do, the Trolley-Preferences Method may build this assumption into the survey questions by stipulating choices (step 2) that human drivers might make. The Trolley-Preferences Method has no mechanism for discovering new options available to AVs. Trolley problems have a use in developing an account of the principles we use to justify certain kinds of harms, but they are not well-designed for developing collision algorithms themselves [19] .

Another final crucial limitation of the Trolley-Preference Method is that public preferences do not reliably track what we, or AVs, ought to do. For example, Awad et al. reported that respondents expressed a strong preference for "businessmen" over "large women", who in turn were preferred over "criminals" in determining who should be killed in an unavoidable crash [13] . Earlier work by Bonnefon et al., moreover, described that individuals preferred their car to be selfish and partial towards them, while preferring other cars to be altruistic and self-sacrificing: an observation that shows that preferences sometimes cannot guide action at all [12] . However, much like other models of moral psychology that aggregate individual as a basis for ethical decision making, the Trolley-Preferences method fails to provide data that translates directly to a) how individuals really make moral decisions; nor b) whether some behavior is ethically justified [20]. Preferences are important in guiding decision-making, but they should be settled, reflective preferences that present our best understanding of what ethics requires of us [21], [22].

II. The Data Theories Method

We present a proof of principle method for designing ethical algorithms for AVs that avoids the limitations of the Trolley-Preferences Method. For given scenarios, we model possible trajectories for an AV using naturalistic and simulated data, and calculate expected injuries from historical car accident data [23]–[26]. We use this approach to generate a set of options and expected outcomes for each scenario, and then we consider what different kinds of ethical theories would say about each scenario to gather evidence about the best ethical algorithm for AVs. This method, the Data-Theories Method, takes the form:

1) Scenario pool: Select any ethical choice scenario an AV could face.
2) Scenario analysis: Use all available relevant data, e.g. crash data, to determine (A) the complete set of acts an AV could perform in the scenario and (B) the probability of various consequences of those options.
3) Ethics data: Determine which options would be obligatory, permissible, prohibited, etc. according to different kinds of ethical theories.
4) Algorithm generation: Where plausible ethical theories agree on an option, AVs should be programmed to make this choice. Where ethical theories disagree, take this as evidence about the best ethical algorithm for AVs. AVs following the best ethical algorithm in this scenario will either choose an option that is favored by at least one plausible ethical theory, or will choose an option that is favored by the best theory of how to compromise between different ethical theories.

Our method follows steps 1) through 3), while our results describe step 4) of the method.

A. Scenario Pool

We developed 16 scenarios as vignettes that involving ethical choices an AV has to make, involving a variety of features (Appendix A). These vignettes were styled in the manner of philosophical thought experiments like the trolley problem and its successors, presenting a range of initial possible options. The research team voted on their preferred scenarios for exploration for this study. A final list of eight was chosen, with ties broken by discussion. After determining available data, three scenarios were chosen for the tractability in developing our proof of principle, specifically that they entail calculable physical risks to those involved in the scenario and choices made by the AV:

1) Tailgater, in which an AV is forced to avoid an obstacle while being tailgated by a human driver.
2) Intervention, in which an AV can intervene to save pedestrians from an out-of-control vehicle.
3) Off-Ramp, in which an AV can leave a highway by imposing risk on an entering vehicle, or continue to the next exit.

Each vignette was then refined for the purposes of specifying what the ethical conditions were for each scenario. For clarity, the full text of each vignette is reproduced in the next section alongside the data collected for analysis.

B. Scenario Data

Each of the three scenarios required different strategies to model, using both shared and distinct data sets. Because this is a proof of principle study, several simplifying assumptions were used. Firstly, historical crash data could be used to predict injuries that would result in occupants of an AV experiencing the same acceleration in the same direction. In reality, autonomy will be a feature of new cars, which

are inherently safer by design than older cars, incorporating features to protect its occupants during a crash. Therefore, the injury risks we predict are likely to be conservative. Secondly, for simplicity, we assumed that all vehicles involved in a crash are the same mass. If AVs tend to have a lower curb weight than human-driven cars, risk to occupants of an AV will be greater than we predict. If they have a higher curb weight, then risk will be lower. We assumed that all crashes had a coefficient of restitution of 0.5, which is reasonable but newer cars may incorporate more crumble zones, reducing the resulting acceleration in a crash.

Vehicle crashes often share common data; we pursued our model using the following historical data from the National Automotive Sampling System (NASS) [26]. Change in velocity following impact (delta-V) for each vehicle was calculated with a restitution model that incorporates vehicle masses and pre-impact velocities [27]. Delta-V was used to predict injury and fatality outcomes following a crash based on historical data as described below. For vehicles, the Maximum Abbreviated Injury Score (MAIS) was used as a measure of injury outcome [25]. A MAIS of 3 or above ($MAIS3+$) is at least serious injury, such as an open fracture of the humerus ($MAIS = 3$), or a perforated trachea ($MAIS = 4$). $MAIS3+$ was considered to be associated with a significant ($P > 8\%$)probability of death (Table I). A MAIS of 2 or less was considered at most a moderate injury.

The probability of injury or fatality for a given delta-V was calculated by fitting a logistic regression model to historical crash outcome data obtained from the NASS Crashworthiness Data System (CDS) for the years 2009-2015 [26]. Collision data were stratified by front, rear, side-on crashes, and by the longitudinal and latitudinal delta-V. Where applicable, probability of pedestrian fatality for a given vehicle velocity was calculated by fitting a logistic regression model to historical pedestrian crash data obtained from the NASS General Estimates System (GES) for the years 2011-2015 [26]. For human driven vehicles we selected a reaction time of 3 seconds. This corresponds to data on human reaction times [28], and US state recommendations for following distance [29]. AVs were assumed to react instantaneously.

Tailgater:

> Scenario (Figure 1: A Tailgater (TG) is closely following the AV, and the AV is following the Front Car (FC) in a single lane road. At the start of the scenario, all cars are currently moving at the same velocity of 90 kph, consistent with highway speeds. The scenario starts with FC suddenly braking to a stop. The AV is responsive enough to stop in time to prevent a collision with FC because the AV is following at a safe distance. However, though the AV is responsive enough to avoid a collision with the lead car, TG may not be responsive enough to avoid crashing into the AV. This scenario is further constrained in that the AV cannot swerve out of the way (due to oncoming traffic on the left and a barrier on the right). Intuitively, the AV appears to have two options: it could slam on the brakes and suffer a severe rear end collision, or it could intentionally ram the forward car at a relatively low speed, reducing the speed of collision of the TG with the AV. Given the AV's superior ability to measure speeds and distances, is there another way of managing potential injuries available that may not be available to a human driver?

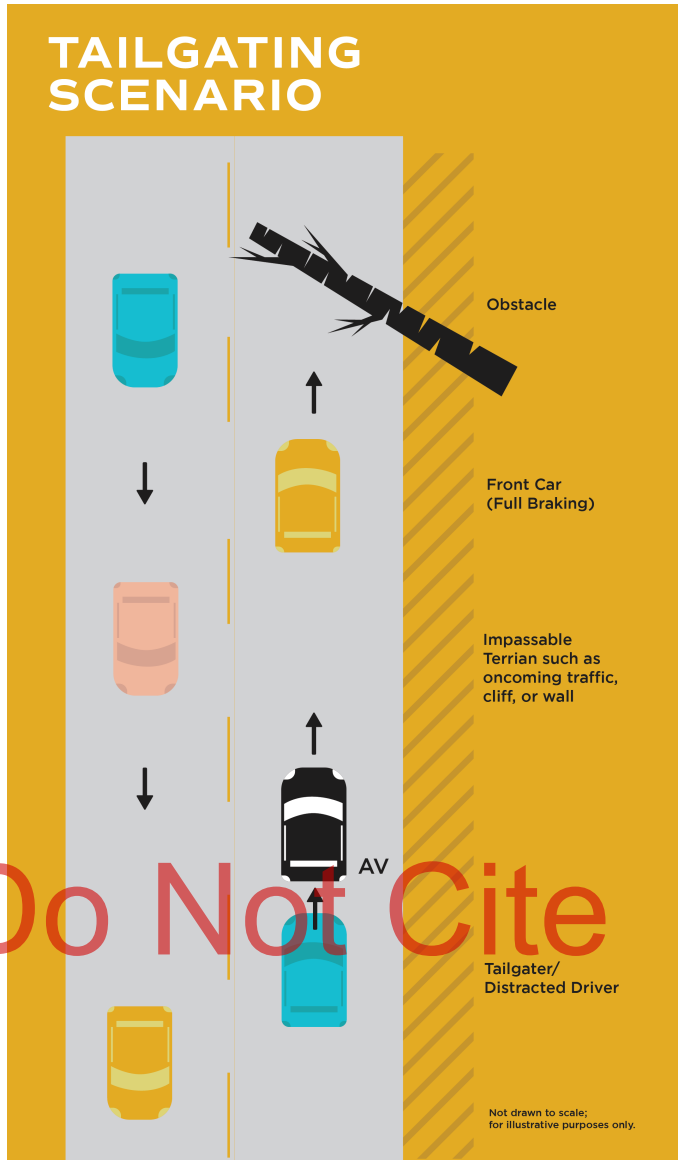

Fig. 1. Tailgating Schematic

Crashes in Tailgating were modeled as two-car collisions, where both cars have the same mass, braking ability, and are travelling in the same plane, resulting in a change in velocity (Δv) [27]:

$$v'_1 = v_1 - (1 - e/2).(v_1 - v_2) \quad (1)$$

$$v'_2 = v_1 + v_2 - v'_1 \quad (2)$$

$$\Delta v_i = v'_i - v_i \quad (3)$$

Where $v_{1,2}$ are the pre-impact velocities of vehicles 1 and 2; $v'_{1,2}$ are the post-impact velocities of vehicles 1 and 2; and e is the coefficient of restitution (determined to be 0.5).

TABLE I
MAIS3+ CODING SCHEME

MAIS-Code	Injury	Example	P(death)
1	Minor	Superficial laceration	0
2	Moderate	Fractured sternum	0.01-0.02
3	Serious	Open fracture of humerus	0.08-0.1
4	Severe	Perforated trachea	0.05-0.5
5	Critical	Ruptured liver with tissue loss	0.05-0.5
6	Maximum	Total severance of aorta	1
9	Not Further Specified		

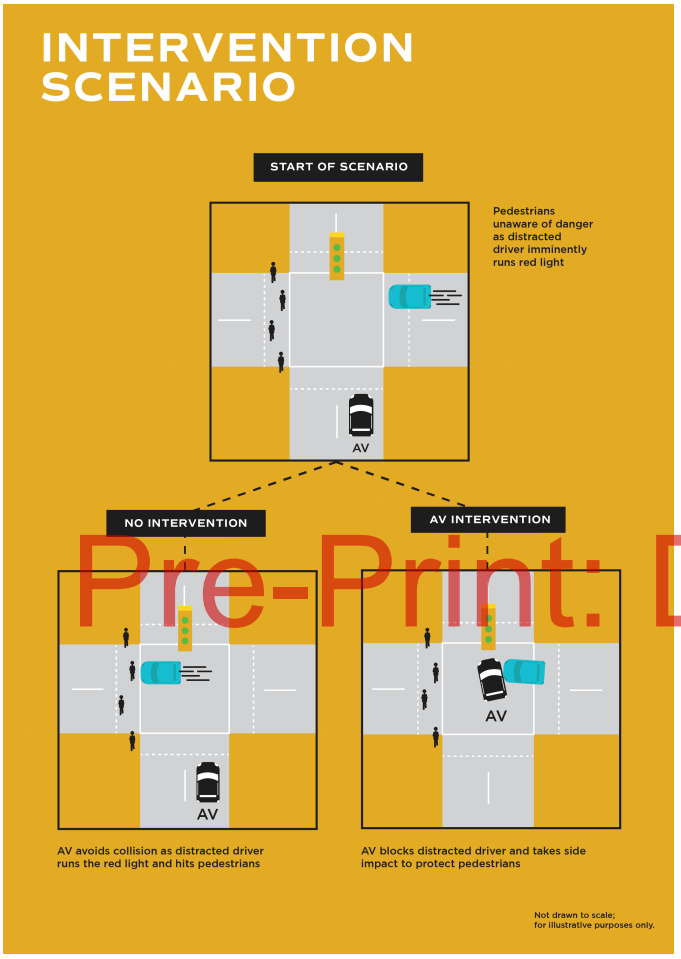

Fig. 2. Intervention Schematic

Intervention:

Scenario (Figure 2: An AV is stopped at an intersection. The light has just turned green, and the AV intends to drive straight through the intersection but detects another car approaching the intersection from the right. The AV correctly infers that the human-driven car (HD) will run the red light and careen into the intersection (because the car is not decelerating). Pedestrians (P) are crossing the crosswalk to the left of the AV, and are in the path of the distracted driver. The AV can stay put, avoiding possible harm to its occupants, but in so doing allows/risks (preventable) harm to the pedestrians. Alternatively, the AV can enter the intersection and force a collision with the distracted driver's vehicle, saving the pedestrians from harm, but in so doing risking harm being done to the occupants of both vehicles.

This model uses the simplified assumption that there are two possible choices:

A: The distracted driver hits a single pedestrian. The pedestrian is the only one at risk of death.

B: The distracted driver hits the AV on the side. The (single) pedestrian has no risk of death, but both the AV and the distracted driver have some risk of death.

We assumed that the AV only has control over whether it is in the path of the distracted driver. We stipulated that other options would be expected to take extra time (e.g. letting the AV occupants out first and then intervening to save the pedestrian(s) on the crosswalk)

Only probability of fatality was used for A because GES uses another standard, the KABCO Injury Classification Scale [30], which is not commensurate with MAIS coding used to report vehicle occupant injuries in the CDS. Vehicle fatality probabilities for front and side crashes used the same methods previously described for Tailgating. Data about side crashes were filtered to include only impacts from the opposite side of the occupant.

Off-Ramp:

Scenario (Figure 3: Risk minimization is generally a good strategy in an AV but can lead to some very frustrating outcomes. Take, for example, a three- or two-part cloverleaf intersection at the junction of two large highways. Traffic is such that an AV attempting to merge onto the other highway cannot do so without assuming perfect behavior on the part of human drivers and/or taking some risks. One of the AV's options is to remain in its lane, leave the highway, then enter back onto the original highway, now heading in the opposite direction. Sometimes there are miles between highway exits. Should an AV extend travel time for its passengers in order to minimize risks? Or should the AV take risks similar to a human driver in a similar situation? (For example, human drivers might "cut off" other drivers, stop in the merge lane, roll onto the road shoulder to stay on the road, etc.) Can the risk of taking a longer and less efficient route be compared to the risk of an accident caused by "reckless" AV behavior?

For this scenario, we decided that a simple dog leg (i.e. a single merge u-turn off the highway) would be sufficient

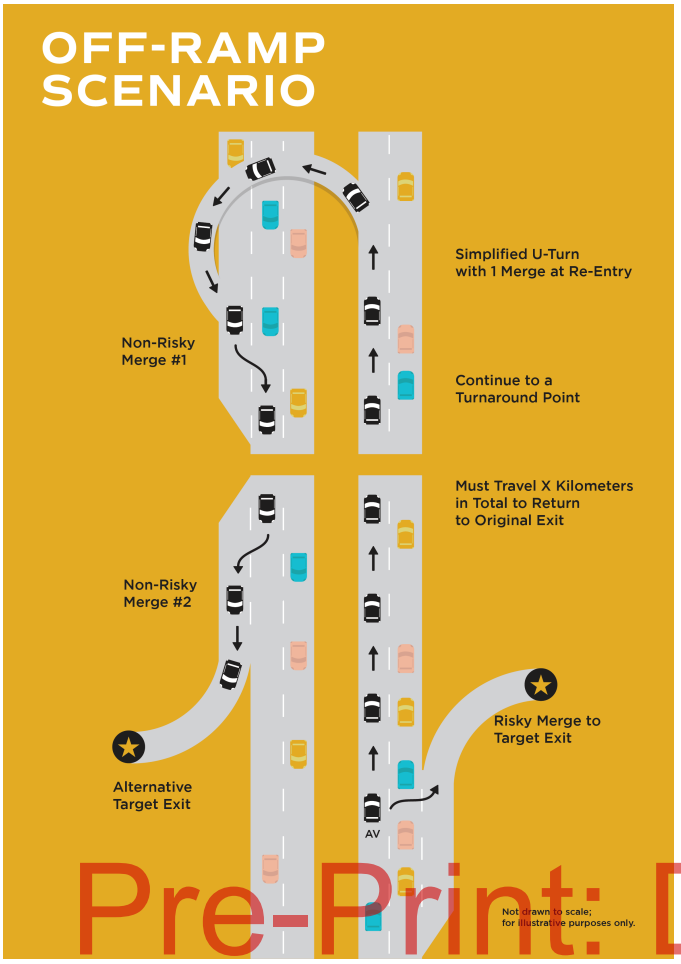

Fig. 3. Off-Ramp Schematic

to probe the problem. This scenario used per-mile accident data from the New York Department of Transportation 2015-2016 Average Accident Rates [31]. An aggressive driving risk modifier was estimated based on results from Habtemichael [32]. The relative risks of continuing for some distance x past a turn and rerouting back to make the original exit as opposed to engaging in a risky exit maneuver is represented as:

$$A_x x + A_n(n_x + 1) = A_n k_e n_N$$

A_x is the rate of accidents per million vehicle miles on a highway, A_n is the rate of accidents per million entrances or exits from the highway, n_X is the number of extra (not "recklessly" risky) merges required to get back on course plus the next safe merge off the highway, k_e is the risk multiplier for a risky merge, and n_e is the number of "reckless" merges attempted (here, we set this value to 1). Therefore, the extra distance x (including merges to retrace one's steps) an AV can go before it incurs more risk than making the risky exit maneuver is:

$$x = A_n(k_e - (n_x + 1))/A_x$$

C. Ethics Data

Ethical analysis was performed through a deliberative process between the philosophers and empirical researchers on the research team. This process was not designed to elicit the preferences of the team, but rather examine each scenario using the available philosophical literature on the ethics of acting under conditions of risk and uncertainty. Our method remains agnostic as to whether AV algorithms should ultimately be "top-down" (e.g. governed by formal methods), "bottom-up" (e.g. using machine learning to create appropriate action), or a hybrid of the two [33], [34].

Because there are many ethical theories, and none are held as true a priori, we chose one important feature of ethical analysis: how ethical theories deal with partiality, or duties to oneself as distinct from others [35]. Some ethical theories are partial and maximizing, i.e. you must always choose the option that maximizes your own welfare. A paradigmatic example of partial maximizing would be moral egoism, according to which an agent's actions are right if and only if they increase that agent's well-being [36]. Impartial Maximizing Approaches, conversely, require an agent to choose the option that maximizes total well-being. An indicative example of impartial maximizing is classical act-utilitarianism, according to which any action is right if and only if it best promotes the aggregate well-being of the world [37].

Many contemporary moral theories often belong somewhere between these two extremes, and acknowledge some limited duty to others [38]–[41]. A common feature among these theories is they recognize that we may be permitted to act in our own self-interest, but have some limited but no less overriding duty to protect or help others even if doing so imposes risk on ourselves.

The use of partiality as our topic of ethical analysis allowed for a high-level analysis of the ethics of AV behavior in each case, in a way that captures a wide range of common ethical theories, without necessarily committing to a particular theory in particular. Agnosticism about which theory is ultimately correct allowed us to map elements of convergence in moral theories that might otherwise be opposed in other important ways.

III. Results

A. Tailgater

In the simple formulation of the Tailgating case, the AV either stops immediately (which harms the TG and AV), or else stops more slowly (in which case harm to TG and AV decreases with the rate at which the AV slows) and potentially collides with the FC. Since front collisions are safer than back collisions — even a low speed, front-end collision (with an instantaneous change in velocity (delta-V) of 10-15 kph) may still result in serious injury (MAIS 3+) (Figure 4)— these two options pit self-interest against the interests of all parties to the accident. Development of our model, however, demonstrated that in all cases an AV can tune a collision between front and back collisions, so that even "selfish" AVs have a range of possible solutions for a range of velocities.

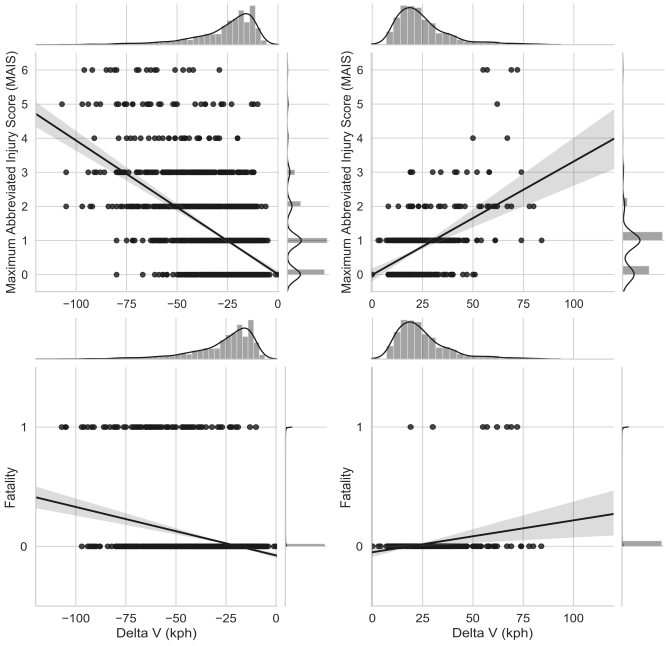

Fig. 4. MAIS (top-row) and fatality (bottom-row) outcomes per vehicle occupant for front (left-column, negative Delta-V) and rear (right-column, positive Delta-V) end collisions from historical crash data (CDS 2009-2015). Dots represent the outcome for individual occupants and the trendline demonstrates an increase in severity with an increasing —Delta-V—. n(front) = 2052, n(rear) = 499

TABLE II
TAILGATING CASES MAIS3+ INJURY PROBABILITIES

Case	Probability of a MAIS3+				Total expected
	Front	AV	TG	Total	MAIS 3+
Case A	0%	0%	0%	0%	0.00
Case B	7%	20%	9%	33%	0.36
Case C	13%	36%	23%	57%	0.72
Case D	0%	2%	4%	6%	0.06
Case E	0%	4%	6%	9%	0.10

Five cases are shown to demonstrate how AV's choices about tuning deceleration decisions can affect outcomes (Figure 5). The scenario is initiated by the front car (in green) braking to a stop due to some object blocking the road, forcing the AV (in pink) and TG (in purple) to react. Injury (Table II) and fatality (Table III) outcome probabilities were estimated for each case.

Case A is the ideal situation, where no crash occurs. The AV tries to give TG as much time to brake as possible, and TG reacts in time (perhaps because the FC is large and can be seen through the AV by the TG). In cases B and C, the AV follows the same procedure, but TG either reacts too slowly or not at all, resulting in crashes. If the AV assumes that TG will not react unless "woken up" with a small impact, it could nudge TG as in Case D. Finally, if the AV assumes that TG will not react at all, it could slow/stop TG by coming in contact and functioning as an "assistive brake" for both vehicles, as in Case E. This scenario is enabled by the fact that the AV has a super-human ability to judge the relative speed of all vehicles around it. The resulting injury and fatality outcomes demonstrate that the AV has the ability to avoid outcomes with

TABLE III
TAILGATING CASES FATALITY PROBABILITIES

Case	Probability of a fatality				E(fatal)
	Front	AV	TG	Total	
Case A	0%	0%	0%	0%	0.00
Case B	4%	7%	2%	13%	0.13
Case C	6%	15%	6%	25%	0.27
Case D	0%	2%	1%	3%	0.03
Case E	0%	3%	2%	4%	0.04

a higher chance of severe harm by choosing outcomes with a higher chance of less severe harm.

B. Intervention

In Intervention, only two options are available to the AV: intervene or don't intervene. Because the collision velocity is perpendicular to the AV's motion, the speed at which the AV intervenes is not a determinant of the result. Rather, the dominant feature of this scenario is the velocity of the human-driven vehicle, which determines the velocity of impact for either the AV or the pedestrian.

Results for both scenarios can be found in Figure 6. In cases where the AV does not intervene, the probability of death for the pedestrian increases rapidly with collision velocity. In cases where the AV does intervene, the probability of death for both the passenger in AV and the human driver of the out-of-control vehicle increases with speed. However, both of these latter probabilities are lower than the probability of death for the pedestrian for any given velocity. Moreover, the likelihood of either the AV passenger or out-of-control driver (or both) dying, taken together, is lower than the probability of the pedestrian dying at any given velocity. The likelihood of the AV driver dying depends on the side of the car being hit. At velocities under 120 kph (which is quite high for speeds observed near four-way intersections), the driver of the AV is more likely to die than the distracted driver.

It may seem that the Partial and Impartial Maximizing approaches are in complete opposition here about what should be done. Since intervening increases the risk to the AV's occupants, it is not required and may even be forbidden according to the Partial Maximizing Approach. Intervening decreases the risk to the pedestrian(s) more than it increases risk to the vehicle occupants, and thus is always obligatory according to the Impartial Maximizing Approach.

Yet our results suggest that the risk to the AV occupants is low so long as the approaching car is moving slowly (up to 60 kph), in contrast to a typical Trolley-Preferences methods cases would stipulate one must choose between a guaranteed driver's death or a guaranteed pedestrian's death. At approximately 40 kph, however, the risk to pedestrians rises quickly. The differences in risk between these intervals suggests that intervening may not required by the Impartial Maximizing Approach unless the distracted driver is moving faster than 40 kph. Further, above some speed The Limited Duty to Others Approach may say that there is some threshold at which point intervening is not required. Perhaps at 60 kph the risk to the AV's occupants becomes too great, or the

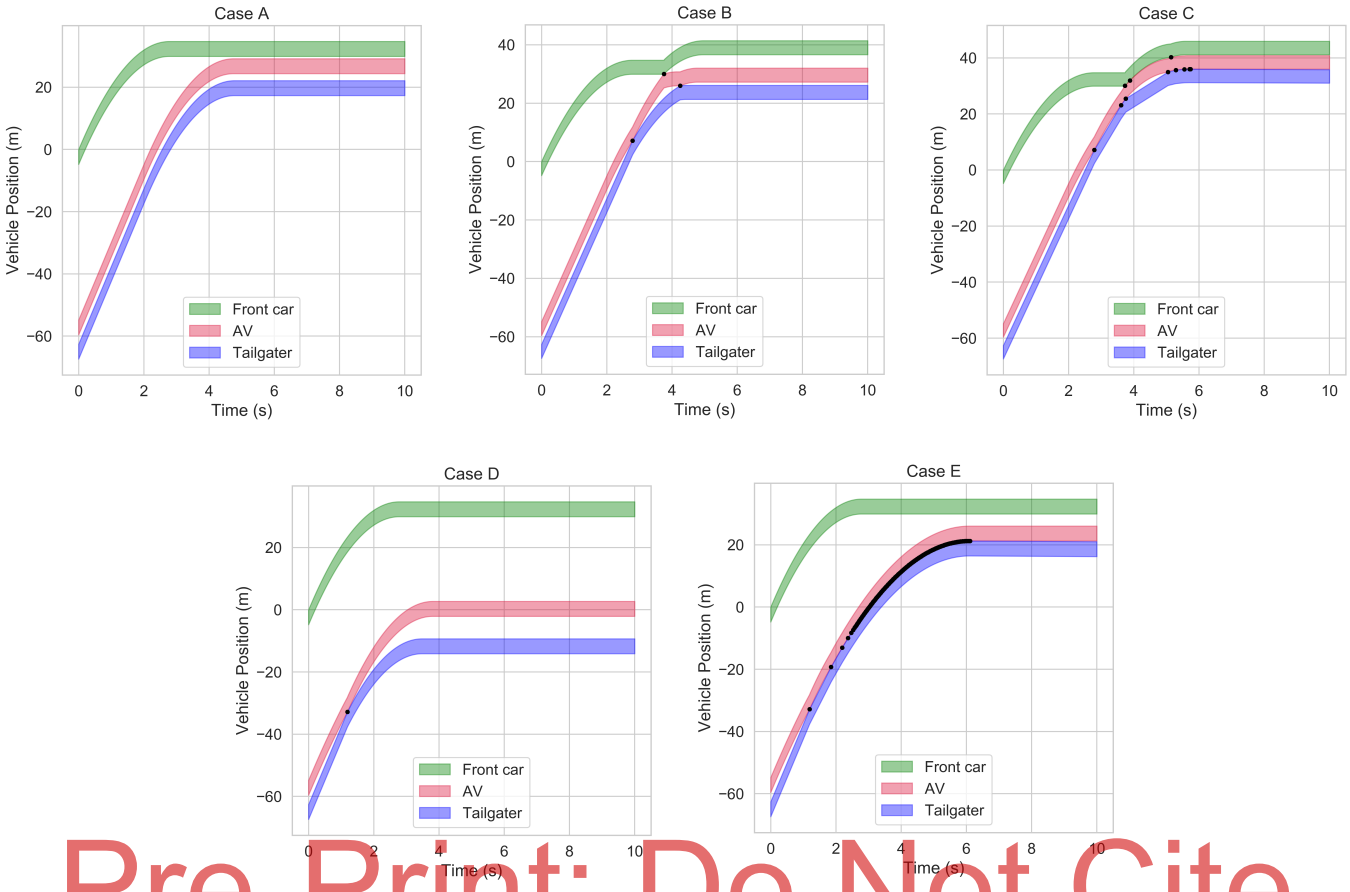

Fig. 5. Tailgating Cases. The position of each vehicle is shown by time and black dots indicate a collision. Case A: The AV brakes as late as possible to allow the longest window for TG react. TG takes 2 seconds to react. Case B: The AV follows the same procedure as Case A. TG takes 3 seconds to react. Case C: The AV follows the same procedure as Case A/B. TG does not react at all. Case D: The AV starts a slower brake sooner to "wake up" TG, who reacts on contact. Case E: TG does not react at all, and the AV acts as a brake assist to slow/stop TG.

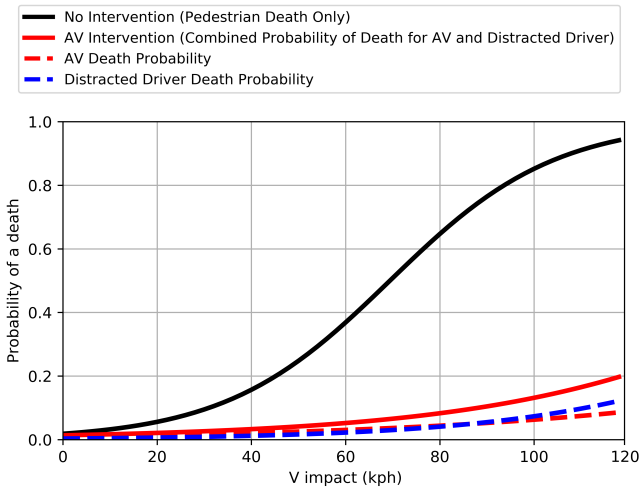

Fig. 6. Pedestrian and vehicle fatality probabilities as a function of velocity of impact. Deaths suffered in No Intervention (black line) are suffered by the pedestrian. In AV Intervention (solid red line), deaths are suffered by vehicle occupants, which are also individually shown for the AV (dashed red line) and distracted vehicle (dashed blue line).

expected total gain in well-being is not worth the risk to the AV's occupants.

Finally, if we loosen the assumption that HD will always hit P in the absence of intervention, the risk of death to P may substantially decrease, as the curve in Figure 6 would represent only the conditional probability of death if a collision occurs. In this case, the minimum speed that obligates a response from the AV may converge with the maximum speed at which an individual is not obligated to accept a risk of serious injury. This may lead to a convergence between Limited Duty to Others and Partial Maximizing Approaches, though showing this would require further investigation.

C. *Off-Ramp*

The extra distance (including merges to retrace one's steps) that an AV can go before it does something as risky as the risky exit maneuver is:

$$x = A_n(k_e - (n_x + 1))/A_x$$

Our extracted data (Table IV) provided indicative values. For a simply dog-leg, $n_x = 1$. For these values, $x = 1.2$ km [0.267km, 1.92km], which corresponds to the additional

distance on the highway to achieve parity with the risk imposed, to all parties, by a risky merge. Our analysis indicates that choosing a risky merge will almost always do more to minimize risk than continuing to drive on the highway in order to loop back around. Driving more than two extra kilometers would almost certainly be more risky than an aggressive merge. The merge was always safer, $x \leq 0$, when $k_e = n_x + 1$.

TABLE IV
VARIABLES IN OFF-RAMP

Variable	Range	Chosen values
A_x	2.0-3.0	2.0
A_n	0.1-0.6	0.5
N_x	1-inf	1

Crashes at merges tended to result in lower fatality rates than crashes on highways: 0.19% of urban non-intersection highway accidents result in fatalities, compared to 0.08% of urban highway off-ramp accidents. To compare the severity of these types of crashes, we used the mean cost of accidents for merges against highway driving, $31,000 and $38,200 per crash respectively. When allowing for the severity of crashes x = 0.976km [0.208, 1.55]. That is, the expected additional distance to achieve parity in risk was reduced by approximately 20% when controlling for the relative cost of the crashes.

Off-Ramp at first appears to be counter-intuitive from the Impartial Maximizing Approach. Arguably, many would think that if merging to get off of the highway would require a risky maneuver, then perhaps the Impartial Maximizing Approach would have the AV take its occupants on a long detour in order to merge safely and reduce overall risk. It's not obvious that the Partial Maximizing Approach would disagree, so long as the detour would reduce risk to the AV's occupants enough to outweigh the inconvenience of a longer travel time, then it would also have the AV take the longer route. But Off-Ramp initially looks like a scenario in which risk-adverse AVs may make choices that their human occupants would be unhappy with.

Our results demonstrate both an interesting example of why AVs might wish to be risk-taking in the short term to trade off risks in the long term, and demonstrate a potential problem familiar to recent social science work on autonomous vehicles that considers framing effects around risk. That is, despite intuitions to the contrary, an AV designed to reduce risk for all parties (in accordance with the Impartial Maximizing Approach) would not be expected to take long detours. Given our assumptions, a detour longer than 1.2 km, or 0.976 km if we account for severity of crashes, would be riskier than an aggressive merge and so morally unwarranted. While this analysis does not reveal a new option, it does show that the scenario is not one in which ethical theories will widely diverge in principle about what should be done. It shows, however, that the action on which ethical theories converge may not readily discernible by the Trolley-Preferences method, as it is not implausible that individuals responding to a similar case in a Trolley-Preferences survey would view an aggressive merge as riskier than a short detour.

IV. DISCUSSION

Our method offers a proof of principle about how philosophers and empirical researchers might collaborate to develop a robust empirical account of ethical algorithms in autonomy vehicle. This method demonstrates how ethical theory might be leveraged to examine empirical data, and provide an account of why AVs ought to make certain decisions in emergent or, in the case of Off-Ramp, prosaic scenarios. This provides an account of ethical AVs that is arguably defensible from the perspective of a variety of stakeholders, and provides a normative basis for responding to certain facts about the road.

This method is not confined to fully autonomous, or "Level 5" AVs. While some of these cases, such as Intervention, may require full or near-full autonomy to enact, level 2 AVs that only perform limited autonomous functions may plausibly be programmed to act in the way we suggest in Tailgater. The requirements for that scenario are fairly straightforward, and require only that the vehicle can recognize cars both in front and behind it, and control its acceleration. This is within the realm of possibility of modern vehicles that have a combination of assisted emergency brake using a rear camera, and existing adaptive cruise control attempts.

Likewise, lane-switching is plausible near-future autonomous vehicles capable of lane shifting and highway entry and exit. These data, and the method in general, do not require that AVs are fully autonomous, though they are useful in those cases as well. With limited adjustments, this method would be useful for developing and implementing existing or emerging technologies into the next generation of vehicles.

Further, our method has clear advantages over previous attempts to empirically determine desirable or ethically justified qualities of AVs. Our method does not rely on spurious connections between individual preferences and ethical principles. Moreover, our method is able to take existing data and, with relatively few assumptions, infer plausible capacities that AVs may have in the future. This gives a platform on which to develop and account of ethical algorithms in which what an AV ought to do is straightforwardly connected to what an AV can, or is believed to be able to do.

V. LIMITATIONS

The limitations with this method are first in data collection, and second in the use of moral theory. Some of our data, for example, requires assumptions to utilize. Intervention presumes that the pedestrian will be struck by the distracted driver. In reality, the pedestrian could notice the oncoming car and dive out of the way, or the distracted driver could swerve at the last moment. Variations on this scenario could include the presence of a large number of people in the intersection, increasing the probability that not only one person would be struck, but multiple people. In Intervention, moreover, there may be small adjustments an AV could make to its trajectory to further minimize risk to the driver (for example, accelerating fast enough so that HD collides with an unoccupied section of AV. Reanalysis of Intervention incorporating a probabilistic model of pedestrian death may also change our conclusions about what the AV ought to do at different velocities. This

uncertainty is important to consider but is impossible to characterize a priori with existing data, and so remains for future work.

A general philosophical limitation of our method is that it doesn't attempt to resolve conflicts between comprehensive ethical theories. Where ethical theories disagree about what an AV should do, our method can discover this, perhaps even shedding light on why they so diverge, but it has no mechanism for determining what should in fact be done. Here we run up against the limits of current ethical and metaethical theory. This mechanism of conflict resolution will almost certainly be required by any project to design ethical algorithms for AVs. That the Trolley-Preferences method doesn't need one in virtue of relying on aggregate, unreflective preferences, is no real advantage—indeed it is a sign that it is unhelpfully off-track.

Our cases, moreover, do not obviate the need for careful analysis of AV algorithms even in cases where there is initial agreement. Consider, for example, a "hard case" version of Tailgater, for example, would loosen the stipulation that the scenario occurs on a single lane road. If, for example, the AV could swerve out of the way of FC, we might ask whether it is obligated to issue a "wake up call" first, given that the AV might avoid FC but TG could fail to do so. And we might ask whether, if the AV has the option to swerve out of the way, it nevertheless ought to act as an emergency brake given the low risks entailed. Here, Partial Maximizing approaches diverge from Impartial Maximizing approaches, and plausibly also Limited Duty to Others Approaches.

This divergence might also arise if, for example, FC is a much lighter vehicle, such as a cyclist. In this case, the risk to AV of a lethal front-end collision is very close to zero, but the risk of a lethal rear end to FC (in this case, being run over by the AV) is very high. Even if the AV were able to swerve, it might be impermissible to do so as it would result in FC's death by being run over by TG. As such, a wake-up call or emergency brake maneuver might be not only permissible but obligatory under a Limited Duty to Others or Impartial Maximizing approach, even though it places the occupant of AV in (a small amount of) harm's way.

Certain vehicles, moreover, may have important duties that regular commuters do not. In the case of Intervention, for example, emergency vehicles (particularly those that are state-owned) may have the obligation to place their passengers in harm's way. We already expect fire fighters, for example, to accept risks on behalf of public safety. We could plausibly require fire fighters in autonomous fire trucks to accept new risks.

VI. Conclusion

The Data-Theories method can help identify new options for AVs (as it does in Tailgating) and can provide new information to use in evaluating an AVs options (as it does in all three scenarios). The Data-Theories method us also more likely than the Trolley-Preferences method to produce an ethical algorithm for AVs that actually has them act ethically as opposed one that has them act in a way that's merely perceived by the public to be ethical.

In this paper we've presented a new method for finding ethical algorithms for AVs. We've applied it to three different scenarios in order to show how it works, what its advantages and disadvantages are, and how it might be extended. This method is also generalizable. It could be applied to many risky driving scenarios to determine how AVs should behave and to ensure that future motorists feel comfortable sharing the road with AVs. Future studies involving this method should use data on specific AVs and the masses (and crash-worthiness) of other cars on the road today.

References

[1] N. J. Goodall, "Can you program ethics into a self-driving car?" *IEEE Spectrum*, vol. 53, no. 6, pp. 28–58, 2016.

[2] US Department of Transportation, *Federal Automated Vehicles Policy - September 2016*, 2016.

[3] P. Lin, "Why ethics matters for autonomous cars," in *Autonomes Fahren*. Springer Berlin Heidelberg, 2015, pp. 69–85.

[4] N. J. Goodall, "Away from trolley problems and toward risk management," *Applied Artificial Intelligence*, vol. 30, no. 8, pp. 810–821, 2016-11, publisher: Taylor & Francis.

[5] L. T. Bergmann, L. Schlicht, C. Meixner, P. König, G. Pipa, S. Boshammer, and A. Stephan, "Autonomous vehicles require socio-political acceptance—an empirical and philosophical perspective on the problem of moral decision making," *Front. Behav. Neurosci.*, vol. 12, 2018, publisher: Frontiers.

[6] B. Meder, N. Fleischhut, N.-C. Krumnau, and M. R. Waldmann, "How should autonomous cars drive? a preference for defaults in moral judgments under risk and uncertainty," *Risk Analysis*, vol. 39, no. 2, pp. 295–314, 2019.

[7] J. Rhim, G. Lee, and J. Lee, "Human moral reasoning types in autonomous vehicle moral dilemma: A cross-cultural comparison of korea and canada," *Computers in Human Behavior*, vol. 102, pp. 39–56, 2020.

[8] P. Liu, R. Yang, and Z. Xu, "How safe is safe enough for self-driving vehicles?" *Risk Analysis*, vol. 39, no. 2, pp. 315–325, 2019.

[9] L. Montoro, S. A. Useche, F. Alonso, I. Lijarcio, P. Bosó-Seguí, and A. Martí-Belda, "Perceived safety and attributed value as predictors of the intention to use autonomous vehicles: A national study with spanish drivers," *Safety Science*, vol. 120, pp. 865–876, 2019.

[10] Noa Kallioinen, Maria Pershina, Jannik Zeiser, Farbod Nosrat Nezami, Gordon Pipa, Achim Stephan, and Peter König, "Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives," *Frontiers in Psychology*, vol. 10, 2019.

[11] Frison, A.K., Wintersberger, P., Riener, A., and Schartmüller, C., "Moral behavior of automated vehicles: The impact on product perception," in *Mensch und Computer 2018 - Workshopband*, R. Dachselt and Wever, G., Eds., 2018.

[12] J. F. Bonnefon, A. Shariff, and I. Rahwan, "The social dilemma of autonomous vehicles," *Science*, vol. 352, no. 6293, pp. 1573–1576, 2016.

[13] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment." *Nature*, vol. 563, no. 7729, pp. 59–64, 2018-11.

[14] P. Foot, "The Problem of Abortion and the Doctrine of the Double Effect," in *Virtues and Vices and other Essays in Moral*. New York: Oxford University Press, 1993, pp. 19–32.

[15] BVMI, "Ethics commission: Autonomated and connected driving," 2017.

[16] S. Nyholm and J. Smids, "The ethics of accident-algorithms for self-driving cars: an applied trolley problem?" *Ethical Theory and Moral Practice*, vol. 19, no. 5, pp. 1275–1289, 2016-07, publisher: Springer Netherlands.

[17] J. Himmelreich, "Never mind the trolley: The ethics of autonomous vehicles in mundane situations," *Ethical Theory and Moral Practice*, vol. 21, no. 3, pp. 669–684, 2018-05, publisher: Springer Netherlands.

[18] M. Raue, L. A. D'Ambrosio, C. Ward, C. Lee, C. Jacquillat, and J. F. Coughlin, "The influence of feelings while driving regular cars on the perception and acceptance of self-driving cars," *Risk Analysis*, vol. 39, no. 2, pp. 358–374, 2019.

[19] G. Keeling, "Why trolley problems matter for the ethics of automated vehicles," *Sci Eng Ethics*, vol. 26, no. 1, pp. 293–307, 2020.

[20] J. Kennett and C. Fine, "Will the real moral judgment please stand up? the implications of social intuitionist models of cognition for meta-ethics and moral psychology," *Ethical Theory and Moral Practice*, vol. 12, no. 1, pp. 77–96, 2009, publisher: Springer.

[21] P. Kitcher, *Science, Truth, and Democracy*. Oxford University Press, 2003-11.

[22] ——, "Philosophy inside out," *Metaphilosophy*, vol. 42, no. 3, pp. 248–260, 2011-04.

[23] M. H. Hosseinlou, H. Ahadi, and V. Hematian, "A study of the minimum safe stopping distance between vehicles in terms of braking systems, weather and pavement conditions," *Indian Journal of Science and Technology*, vol. 5, no. 10, pp. 3421–3427, 2012-10.

[24] A. Laureshyn, T. De Ceunynck, C. Karlsson, \. Svensson, and S. Daniels, "In search of the severity dimension of traffic events: Extended delta-v as a traffic conflict indicator," *Accident Analysis & Prevention*, vol. 98, pp. 46–56, 2017-01, publisher: Pergamon.

[25] T. A. Gennarelli and E. Wodzin, "AIS 2005: a contemporary injury scale," *Injury*, vol. 37, no. 12, pp. 1083–1091, 2006-12.

[26] NHSTA. (2016) National automotive sampling system (NASS). Last Modified: 2020-03-05T11:27-05:00 Library Catalog: www.nhtsa.gov.

[27] M. Batista and G. Zovak, "A resistution model of two-car collinear collisions," *Promet - Traffic & Transportation*, vol. 19, no. 1, pp. 1–6, 2007.

[28] D. V. McGehee, E. N. Mazzae, and G. H. S. Baldwin, "Driver reaction time in crash avoidance research: Validation of a driving simulator study on a test track:," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 44, no. 20, pp. 320–323, 2016, publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[29] Maryland DOT, *Maryland Driver's Manual*. Maryland DOT, 2018.

[30] US Department of Transportation. (2019) KABCO injury classification scale and definitions.

[31] NY DOT. (2016) Average accident rates. [Online]. Available: https://www.dot.ny.gov/divisions/operating/osss/highway-repository/Average%20Accidents%20Rates%20Table_2016.pdf

[32] F. G. Habtemichael and L. de Picado Santos, "Crash risk evaluation of aggressive driving on motorways: Microscopic traffic simulation approach," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 23, pp. 101–112, 2014.

[33] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics and Information Technology*, vol. 7, no. 3, pp. 149–155, 2005.

[34] C. Allen and W. Wallach, *Moral Machines*. Oxford University Press, 2009.

[35] S. Scheffler, *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford University Press, 1994-08-11, publication Title: The Rejection of Consequentialism.

[36] K. Burgess-Jackson, "Taking egoism seriously," *Ethical Theory and Moral Practice*, vol. 16, no. 3, pp. 529–542, 2012-06, publisher: Springer Netherlands.

[37] J. S. Mill, *Utilitarianism*. Oxford University Press, 1861.

[38] J. Kleinig and N. G. Evans, "HUMAN FLOURISHING, HUMAN DIGNITY, AND HUMAN RIGHTS," *Law and Philosophy*, vol. 32, no. 5, pp. 539–564, 2013-09, publisher: Springer.

[39] I. Kant, *Grounding for the Metaphysics of Morals*, ser. with On a Supposed Right to Lie because of Philanthropic Concerns. Hackett Publishing, 1993-06.

[40] D. R. Mapel, "Moral liability to defensive killing and symmetrical self-defense," *Journal of Political Philosophy*, vol. 18, no. 2, pp. 198–217, 2010-06, publisher: John Wiley & Sons, Ltd (10.1111). [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9760.2009.00340.x

[41] G. Keeling, "Legal necessity, pareto efficiency & justified killing in autonomous vehicle collisions," *Ethic Theory Moral Prac*, vol. 21, no. 2, pp. 413–427, 2018.

APPENDIX

A. Initial Vignettes (Asterisks denote cases used in analysis)

Tailgating*: An AV is closely following another car. In response to some event, this leading car slams on its brakes. The AV is responsive enough to stop in time to prevent a collision with the lead car. However, the AV is being tailgated by a human driver, and though the AV is responsive enough to avoid a collision with the lead car, the tailgater is not responsive enough to avoid crashing into the AV.

Crosswalk: An AV is sitting at a red light at an intersection, waiting to go straight. It detects a large vehicle, e.g. a semi truck, bearing down on it from behind. The AV infers that the semi will not be able to come to a complete stop before it collides with the AV. The AV has two options: (1) the AV can stay put (or perhaps "dig in" by engaging its parking brake); or (2) the AV can accelerate into the intersection to dampen the impending collision with the semi truck. The AV detects that there are pedestrians crossing in the crosswalk, who risk being injured if the AV moves forward.

Intervention*: An AV is stopped at the front of an intersection with a just-turned green light, and intends to go through the intersection. It detects in its periphery another car approaching the intersection from the right rapidly — suppose it is operated by a human driver who is distracted. The AV correctly infers that the human-driven car will run the red light and careen into the intersection. Pedestrians have begun crossing in the parallel crosswalk to the left of the AV, and are in the path of the distracted driver. The AV can either stay put and avoid harm, but will likely witness harm to the pedestrians, or can enter the intersection and force a collision with the distracted driver, saving the pedestrians but harming the occupants of both vehicles. The risk of death of any occupants is slight, but present. In this scenario, the AV has the choice of avoiding all harm to the occupants but, via inaction, creating harm for others, or causing some harm to the occupants to prevent grave harm to others.

Brake Failure: An AV crests one of San Francisco's famous hills, and begins descending, only to find something sharp piece of metal driven over while headed uphill has cut the line from the primary master brake cylinder and the AV has lost all hydraulic braking. There is no oncoming traffic, but at the next intersection several pedestrians are crossing in front of the AV. The AV can still steer. Should the AV crash into (some) pedestrians—or at least subject them to risk by steering into them—or steer off the road into a heavy object such as a parked car in an attempt to stop, risking the driver?

Chicken: An AV is traveling along a narrow road in autonomous mode. It encounters another car coming from the other direction. The other car is operated by a group of mischievous teenagers who recognize the make and model of the AV, identify it as autonomous, and decide to give the occupants of the AV a scare. At the last minute, they swerve into the path of the AV to threaten an imminent collision. The teenagers are playing "chicken," and expect the AV to swerve out of the way, since it has been programmed to avoid collisions. What the AV's programmers did not anticipate is that, in cases like this, avoiding a collision would reduce the amount of overall harm, but concentrate that harm on the innocent party. And, moreover, there really is no way of knowing whether the teenagers would swerve first.

Off-Ramp*: Risk minimization is generally a good strategy in an AV but can lead to some very frustrating outcomes. Take, for example, a three- or two-part cloverleaf intersection at the junction of two large highways. Traffic is such that an AV attempting to merge into the highway cannot do so without assuming perfect behavior on the part of human drivers and/or taking some risks. The option is to remain in its lane and exit with traffic back onto the original highway, heading in the wrong direction. Sometimes there are tens of miles between highway exits. So, the question is: how much risk taking behavior should an AV take on to avoid extra time for the passengers who would be understandably frustrated and mad about the delays? Should the AV take risks similar to a human driver in a similar situation ("cut off" another driver, stop in the merge lane, etc). Can the risk of driving for additional time and mileage be compared to the risk of an accident caused by "reckless" behavior of the AV? (That is, can you compare a 0.01% accident risk caused by driving an additional 50 miles due to missing the merge to an additional 0.01% accident risk caused by cutting off another driver?)

World War AV: This scenario is similar to the "chicken" scenario in that it involves humans recognizing AVs and taking reckless or illegal action that leverages the AV's perceived weaknesses. Once the risk avoidance behaviors of AVs are widely known, criminals could begin to take advantage. In this case, a team of a pedestrian and a trailing car force an AV to stop by the pedestrian stepping into the road in the path of the AV (and the car preventing it from backing up). The mission of the team is to stop the AV to rob its occupants. If the car was driven by a human, at the first sign of danger, the human may chose to escape by running over the pedestrian or take other action that may intentionally cause an accident. Can the AV be programmed with a "danger mode" that could take aggressive action to enable its passengers to escape harm. Should passengers be able to press a "panic button" that allows an AV to act more aggressively to escape? Does coding this possibility in to an AV make them more prone to hacking?

Mind the Gap: Increased proliferation of the Internet of things allows an AV access to the consumer preferences of its passengers, and—much like our preferences are used to inform the content of websites, everything from Facebook to Google Scholar—could be used to determine the kinds of routes our cars take. Combining purchasing data for a consumer in their AV, for example, could be used to choose routes that take them past their favorite coffee shop chain. This kind of behavior can be expressed in terms of value: the additional growth that comes from advertising revenue, potentially higher consumer satisfaction of preferences (though perhaps not their genuine preferences), or even decreased cost to purchase or provide AVs if they are partly subsidized by advertisers.

Say, however, that this route takes a car through a school zone, rather than a less travelled zone of traffic. What is the consequence for including the value that inheres to advertising above, with, for example, the risks of children walking out into traffic?