Risk Analysis, Vol. 39, No. 2, 2019 DOI: 10.1111/risa.13178

How Should Autonomous Cars Drive? A Preference for Defaults in Moral Judgments Under Risk and Uncertainty

Björn Meder,[1,2,*] Nadine Fleischhut,[3] Nina-Carolin Krumnau,[2] and Michael R. Waldmann[4]

Autonomous vehicles (AVs) promise to make traffic safer, but their societal integration poses ethical challenges. What behavior of AVs is morally acceptable in critical traffic situations when consequences are only probabilistically known (a situation of risk) or even unknown (a situation of uncertainty)? How do people retrospectively evaluate the behavior of an AV in situations in which a road user has been harmed? We addressed these questions in two empirical studies ($N = 1{,}638$) that approximated the real-world conditions under which AVs operate by varying the degree of risk and uncertainty of the situation. In Experiment 1, subjects learned that an AV had to decide between staying in the lane or swerving. Each action could lead to a collision with another road user, with some known or unknown likelihood. Subjects' decision preferences and moral judgments varied considerably with specified probabilities under risk, yet less so under uncertainty. The results suggest that staying in the lane and performing an emergency stop is considered a reasonable default, even when this action does not minimize expected loss. Experiment 2 demonstrated that if an AV collided with another road user, subjects' retrospective evaluations of the default action were also more robust against unwanted outcome and hindsight effects than the alternative swerve maneuver. The findings highlight the importance of investigating moral judgments under risk and uncertainty in order to develop policies that are societally acceptable even under critical conditions.

KEY WORDS: Autonomous vehicles; defaults; moral judgment under risk and uncertainty

1. INTRODUCTION

1.1. Background

The development of self-driving autonomous vehicles (AVs) poses both technological and ethical challenges. AVs promise to reduce accidents resulting from driver errors, such as inattention, perceptual errors, and speeding, which account for more than 90% of accidents in the United States (Singh, 2015). However, even a perfectly functioning AV will not be able to avoid every collision because of the dynamics and uncertainties of driving in real-world environments that include AVs, human-operated vehicles, bicyclists, and pedestrians (Goodall, 2014). Situations can and likely will occur in which all

[1]MPRG iSearch, Max Planck Institute for Human Development, Berlin, Germany.
[2]Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany.
[3]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.
[4]Department of Psychology, University of Göttingen, Göttingen, Germany.
*Address correspondence to Björn Meder, MPRG iSearch, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany; meder@mpib-berlin.mpg.de, bmeder@posteo.de.

possible actions of an AV (e.g., emergency braking, evasive maneuvers) will result in a collision with some obstacle in the environment, given the present state of the robotic system and the environmental conditions (Fraichard & Asama, 2004).

What policies should govern the behavior of AVs in critical traffic situations in which other road users could be harmed? What factors shape people's perception of AV behavior in critical situations or when accidents have occurred? And what ethical considerations should constrain the behavior of AVs? These questions have engaged a variety of disciplines, including robotics, transportation research, philosophy, and psychology (Gerdes & Thornton, 2015; Lin, 2015; Nyholm & Smids, 2016). Several researchers have suggested that critical traffic situations might exhibit structural similarities to moral dilemmas extensively studied in philosophy and psychology, such as the trolley dilemma. Moral dilemmas represent situations in which bad outcomes are unavoidable and conflicting moral reasons exist for each available course of action, for instance, when having to choose between two actions that will both lead to harm for one or more people. Originally conceived of as a thought experiment (Foot, 1967), variants of the trolley dilemma have recently been used to investigate how humans morally judge possible behaviors of AVs in critical traffic situations (Bonnefon, Shariff, & Rahwan, 2016; Powell, Cheng, & Waldmann, 2016). For instance, given a description of an accident scenario, people had to decide whether an AV should either kill pedestrians in the street by staying in the lane or sacrifice itself and its passenger by swerving into a barrier (Bonnefon et al., 2016). When others were concerned, subjects of the study endorsed a consequentialist moral view that aimed to minimize expected loss (e.g., number of casualties): they preferred everyone to operate AVs that would sacrifice their passengers for the greater good (e.g., minimizing the total number of people killed). And yet, they objected to their own car behaving this way. If they themselves were in an AV, subjects expressed a preference for AVs that are programmed in a way that preferentially saves their own lives.

1.2. Moral Judgment and Decision Making Under Risk and uncertainty

Critical traffic incidents are bound to happen. Yet, to what extent trolley dilemmas resemble real-world traffic situations is disputed (Nyholm & Smids, 2016), raising doubts about their appropriateness for informing policy making. In a typical trolley dilemma, all relevant factors and decision consequences (e.g., the number of people killed by an action) are known and certain. Such idealized scenarios with complete information are useful to isolate factors determining people's moral intuitions or to pit different moral theories against each other (Waldmann, Nagel, & Wiegmann, 2012). Yet they can be misleading when exploring what AV behaviors people find morally acceptable because AVs have to operate under risk and uncertainty, that is, operate in situations in which decision consequences are only probabilistically known (a situation of risk) or are even unknown or unquantifiable (a situation of uncertainty) (Knight, 1921; Meder, Le Lec, & Osman, 2013). For instance, if an AV detects the presence of a moving obstacle on the road (e.g., a pedestrian or wildlife), a collision likelihood could be computed based on the speed and mass of the vehicle, possible trajectories of the obstacle, and sensor input about environmental conditions. Such estimates, however, might not be available for all possible actions. For instance, an evasive maneuver is a more complex action than staying in the lane and performing an emergency stop. As an engineer working on AVs put it:

> It takes some of the intellectual intrigue out of the [trolley] problem, but the answer is almost always "slam on the brakes".... You're much more confident about things directly in front of you, just because of how the system works, but also your control is much more precise by slamming on the brakes than trying to swerve into anything. So it would need to be a pretty extreme situation before that becomes anything other than the correct answer. (Hern, 2016)

This highlights that not all possible actions offer the same degree of control, which adds additional uncertainty to their outcomes. Trolley dilemmas are typically described as determinate, with a limited set of courses of actions and definite consequences, whereas in real-world situations it may not even be clear what action would minimize casualties. The engineer's quote also suggests that an action such as staying in the lane and performing an emergency stop could be considered a *default rule*: it is standard routine in many traffic situations, conforms to the general rules of driving, and provides a better degree of controllability. For instance, human drivers are commonly advised to stay in the lane and brake when faced with a possible collision with wildlife, rather than swerving, to avoid losing control of their vehicles.

The distinction between situations of certainty (complete knowledge), risk (probabilistic knowledge), and uncertainty (incomplete knowledge with known and unknown unknowns) is critical from both an applied and a theoretical perspective. From a policy point of view, it is important that AVs implement actions that are societally acceptable even in critical situations and after accidents have occurred. However, if there is a mismatch between the scenarios used to investigate what is morally acceptable to the public and the conditions under which AVs operate, behavioral research can provide only limited guidance for the development of accident algorithms deemed acceptable by society. Public acceptability will be particularly critical when AV-related accidents happen. For instance, in 2016 a driver died in a highway collision because he relied on the car's advanced driver-assistance systems, leading to legal and public examination (Boudette, 2017). Certainly, such cases will happen in the future, too, as technology advances and is adopted more widely.

Theoretically, uncertainty is interesting as it impedes a consequentialist analysis that aims to minimize loss because the relevant information—probabilities, decision outcomes, and their values—may not be available. Uncertainty also has implications for other kinds of policies, such as deontic rules that impose hard constraints on behavior (e.g., "An AV should always avoid collisions with another vehicle, except when this would lead to a collision with a pedestrian or cyclist"). Although such rules are intuitively plausible, applying them requires knowing the consequences of alternative actions to ensure no pedestrians or cyclists are endangered. In contrast, default rules such as "Stay in the lane and perform an emergency stop" do not require any consideration of alternatives or likelihoods. Under risk and uncertainty, rules that work well even under limited information may thus be most useful and result in more transparent and societally acceptable behavior.

2. GOALS AND SCOPE

We investigated what AV behaviors people prefer and find morally acceptable in situations involving risk and uncertainty. Experiment 1 investigated trolley-like dilemmas with likelihoods of decision consequences being either probabilistic (Brand & Oaksford, 2015; Shenhav & Greene, 2010) or unknown. We explored if people's moral judgments and decision preferences are sensitive to probability information, how they differ in situations of risk versus uncertainty, and to what extent people endorse a consequentialist approach aiming to minimize expected harm over the use of default rules.

Experiment 2 examined how people morally evaluate AV behavior in retrospect when a collision between an AV and another road user has occurred. Because accidents involving AVs will not always be avoidable, this retrospective moral evaluation of an action is as important as it is problematic: although people should ignore information that was not available at the time of the decision, *hindsight effects* (Hawkins & Hastie, 1990) and *outcome biases* (Baron & Hershey, 1988) are established findings. People tend to believe that events were more predictable once the outcome is known and have difficulty disregarding outcome information in their moral evaluations (Fleischhut, Meder, & Gigerenzer, 2017). Actions that are less susceptible to hindsight and outcome effects should thus be preferable from a policy-making perspective.

3. EXPERIMENT 1

In Experiment 1, an AV must implement one of two options: stay in the lane or swerve. Staying in the lane puts a pedestrian in the street in danger, and swerving puts a bystander on the sidewalk in danger. We varied the likelihood of colliding with the pedestrian and whether the bystander's collision likelihood was specified (a situation of risk) or unknown (a situation of uncertainty). In all conditions we compared moral judgments and decision preferences for AVs and human drivers (Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015; Powell et al., 2016; Voiklis, Kim, Cusimano, & Malle, 2016).

3.1. Subjects and Design

Data were collected online via the Amazon Mechanical Turk (AMT) platform (Berinsky, Huber, & Lenz, 2012; Casler, Bickel, & Hackett, 2013; Levay, Freese, & Druckman, 2016). To ensure sufficient data quality, the task was available only to U.S. residents with a reliable work reputation (i.e., who had completed 95% of their accepted tasks; cf. Peer, Vosgerau, & Acquisti, 2014). Subjects who completed the task were paid $1 USD for 11.2 minutes on average. We used a 3 (pedestrian's collision likelihood: 20% vs. 50% vs. 80%) $\times$ 2 (bystander's collision likelihood: 50% or unknown) $\times$ 2 (human driver vs. AV) between-subjects design. Subjects were randomly assigned to the conditions.

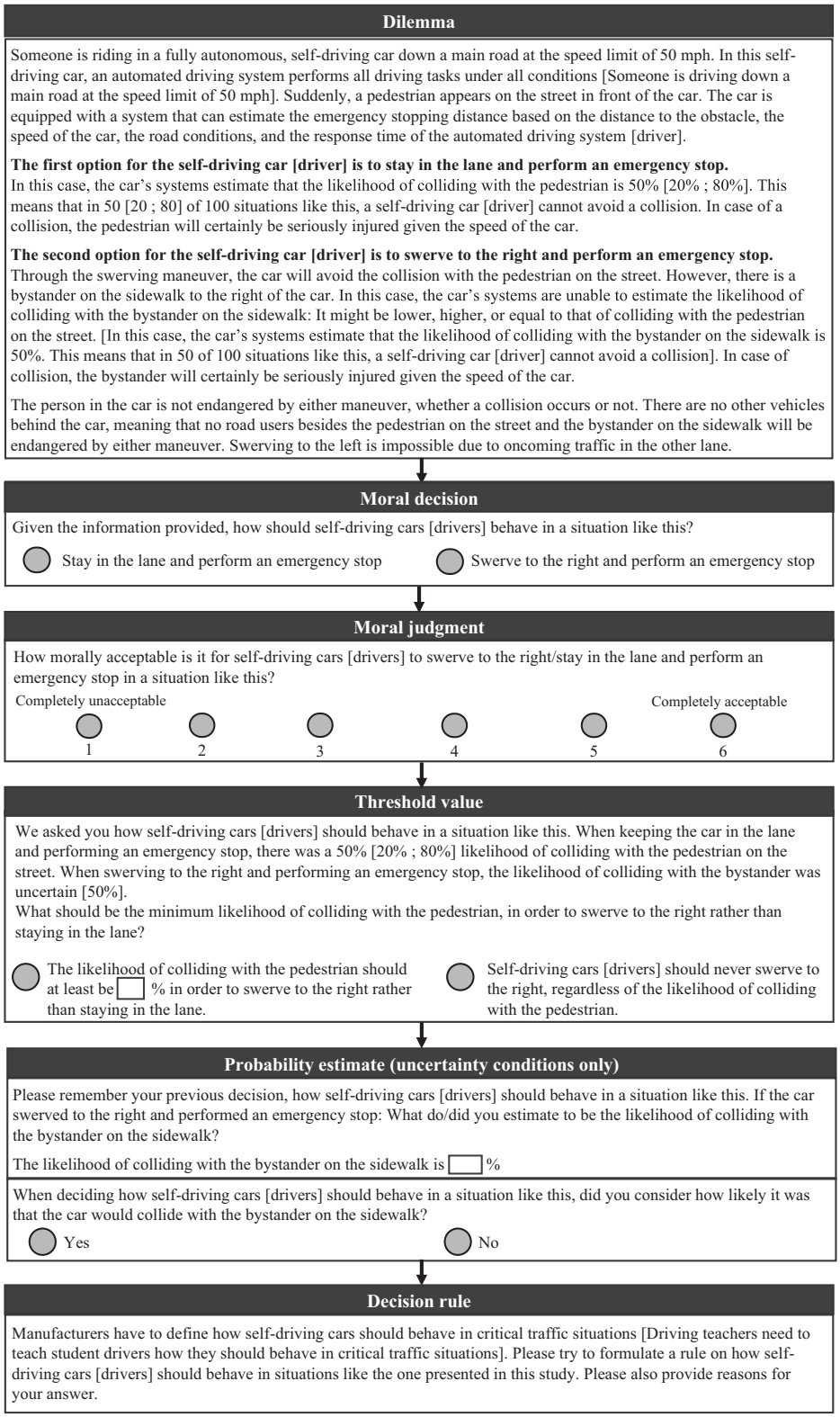

Fig. 1. Overview of experimental procedure and information given to subjects in Experiment 1. Words in brackets indicate variations between conditions; order of answer options was randomized across subjects.

3.2. Procedure and Materials

Subjects were presented with a written description of a traffic situation in which a car is traveling down a road when suddenly a pedestrian appears in its path (Fig. 1). Depending on condition, the car was either an AV with one passenger or steered by a human driver. The first option was to stay in the lane and perform an emergency stop, in which case the car might collide with the pedestrian in the street. The second option was to swerve to the right and perform an emergency stop, in which case the car might collide with a bystander on the sidewalk.

The likelihood of colliding with the pedestrian in the street was either 20%, 50%, or 80%, as estimated by the car's systems. In the *risk condition,* the likelihood of colliding with the bystander on the sidewalk was 50%, as estimated by the car's systems. In the *uncertainty condition* the likelihood of colliding with the bystander was unknown because the car's systems were unable to estimate this likelihood. The possible loss of the two actions was identical: in the event of a collision either the pedestrian or the bystander would be seriously injured but no one else.

The instructions asked subjects to carefully read the scenario because they could only proceed and complete the study after passing an instruction test. The test consisted of six simple multiple-choice questions about the given information (e.g., collision likelihoods for pedestrian and bystander). If a subject failed to correctly answer all questions twice (i.e., even after reading the scenario a second time), the experiment ended.

After the instruction test, subjects were shown the scenario again and answered a series of questions (Fig. 1). The first question asked for a *decision preference*: Should the car stay in the lane or swerve? The second question asked for a *moral judgment*: How morally acceptable is it to stay or to swerve? Each action was judged on a scale of 1 (*completely unacceptable*) to 6 (*completely acceptable*). The third question asked for a *swerving threshold*: What should be the minimum likelihood of colliding with the pedestrian, in order to swerve to the right rather than staying in the lane? Subjects could either answer that self-driving cars [drivers] should never swerve, regardless of the likelihood of colliding with the pedestrian, or provide a numerical estimate between 0% and 100% (in steps of 1%). The fourth question asked for a *probability estimate* of the likelihood of collision with the bystander; this question was asked only in the uncertainty conditions, in which this likelihood was unknown. In addition, subjects were asked to indicate if they considered the likelihood of colliding with the bystander on the sidewalk when deciding how an AV [driver] should behave. The fifth question asked for a *decision rule*; this was an open-ended question in which we asked subjects to come up with a general decision rule for situations like the one described in the scenario. The survey ended after subjects answered some demographic questions.

3.3. Results

To ensure sufficient statistical power, we aimed for at least 60 people per condition, with data collection continuing until enough subjects had completed the study. In total, 1,648 subjects accepted the task on the AMT platform. Of these, 469 did not pass the instruction test and 301 dropped out during the survey. Three subjects were excluded from the analyses because they indicated they had participated before, and three because they indicated insufficient language proficiency. Analyses are based on the remaining $N = 872$ subjects; the number of subjects in the individual conditions ranged from 62 to 82.

3.3.1. How Should AVs and Human Drivers Behave?

Fig. 2 shows the proportion of subjects who preferred to swerve, under risk and uncertainty and as a function of the likelihood of colliding with the pedestrian. For both AVs and human drivers, people's decision preferences varied depending on the likelihood of colliding with the pedestrian and under risk and uncertainty (for results of a Bayesian logistic regression, see Table I). Aggregating over driver type, more subjects opted for swerving under uncertainty than under risk when the likelihood of colliding with the pedestrian in the street was 20%, $\chi^2(1, N = 291) = 26.9$, $p < 0.0001$, or 50%, $\chi^2(1, N = 288) = 11.2$, $p = 0.0008$. Conversely, when the likelihood was 80%, fewer subjects opted for swerving under uncertainty than under risk, $\chi^2(1, N = 293) = 7.45$, $p = 0.006$. Thus, people responded differently to situations of risk and uncertainty, consistent with findings in the decision-making literature (Camerer & Weber, 1992; Ellsberg, 1961; Wakker, 2010).

The decision proportions show a general preference for staying in the lane, even if that was not the loss-minimizing action. When the likelihood of colliding with the pedestrian as well as the bystander was 50%, the expected loss of both actions was equal.

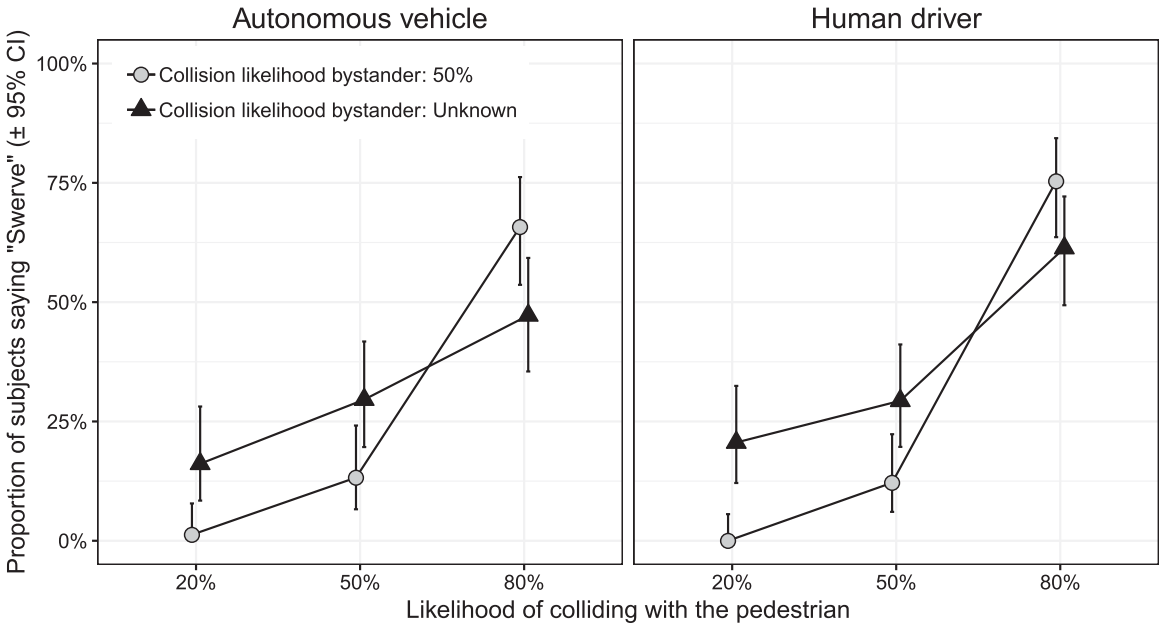

Fig. 2. Experiment 1. Decision preferences regarding whether an autonomous vehicle or human driver should stay in the lane or swerve to the right, under risk (circles) and uncertainty (triangles), as a function of likelihood of collision with the pedestrian in the street.

Table I. Bayesian Logistic Regression Results for Decision Preferences in Experiment 1 (Coded 0 = Stay, 1 = Swerve)

	Coefficient	*SE*	*p*
Intercept	–3.836	0.582	<0.001
Driver	–0.483	0.686	0.482
likelihood pedestrian: 50%	1.911	0.640	0.003
likelihood pedestrian: 80%	4.470	0.618	<0.001
Driver × Likelihood pedestrian: 50%	0.392	0.765	0.608
Driver × Likelihood pedestrian: 80%	0.934	0.733	0.203
Likelihood bystander	2.132	0.625	0.001
Driver × Likelihood bystander	0.784	0.723	0.278
Likelihood pedestrian: 50% × Likelihood bystander	–1.069	0.707	0.130
Likelihood pedestrian: 80% × Likelihood bystander	–2.866	0.688	<0.001
Driver × Likelihood pedestrian: 50% × Likelihood bystander	–0.698	0.840	0.406
Driver × Likelihood pedestrian: 80% × Likelihood bystander	–0.664	0.810	0.413
N			872
McFadden's R^2			0.26

Note: Dummy variable for driver, coded 0 = autonomous vehicle and 1 = human driver. Dummy variable for likelihood pedestrian, coded 0 = 20% and 1 = 50% and 80%, respectively. Dummy variable for likelihood bystander, coded 0 = 50% and 1 = unknown. Estimates based on Bayesian logistic regression using independent Cauchy priors with center 0 and scale 2.5 (Gelman & Su, 2016; Gelman, Jakulin, Pittau, & Su, 2008).

Nevertheless, for both the AV and the human driver more than 85% of subjects showed a preference for staying in the lane (both $p < 0.0001$; here and in the following binomial test against chance). Under uncertainty, when the likelihood of colliding with the bystander was unknown, about 70% of people opted for staying ($p < 0.001$ for both the AV and the human driver).

The preference for staying was also observed when the likelihood of colliding with the pedestrian was 80%. Under risk, when the collision likelihood for the bystander was known to be 50%, a consequentialist policy that assigns equal costs to both collisions entails swerving to minimize expected loss. No one opted for swerving when the likelihood of colliding with the pedestrian was 20% and the likelihood of colliding with the bystander was 50%. Conversely, every consequentialist should opt for swerving when the collision likelihood for the pedestrian was 80% and the likelihood for the bystander was 50%. Yet, in the case of the AV, only 66% of subjects opted for swerving; in the case of a human driver, 75% preferred to swerve. Although both proportions differ from 50% (both $p < 0.01$), clearly not all subjects followed a consequentialist analysis.

Even fewer voted for swerving when the risk for the bystander was unknown. For human drivers, there was still a slight preference for swerving, with

61% choosing this action ($p = 0.06$), but for an AV, subjects were undecided, with only 47% indicating that the car should swerve ($p = 0.72$).

3.3.2. How Morally Acceptable Is Staying or Swerving?

For analyzing judgments of moral acceptability, we computed for each subject the difference between the moral acceptability of staying and swerving. A positive difference indicates that staying in the lane was considered more acceptable and a negative difference indicates that swerving was considered more acceptable. Fig. 3 shows the mean differences; Table II shows the results of an analysis of variance (ANOVA) with driver type, likelihood of colliding with the pedestrian, and likelihood of colliding with the bystander as between-subject factors.

Two key findings were obtained. First, under risk, the relative acceptance of the two options strongly varied as a function of the likelihood of colliding with the pedestrian (Fig. 3, left). Generally, staying in the lane was considered more acceptable than swerving, with the acceptability of staying being even more pronounced for AVs than for human drivers. The more likely the collision with the pedestrian was, the less acceptable it became to stay in the lane and the more acceptable it became to swerve, consistent with the decision preferences. Note that even when the likelihood for both collisions was 50%, staying was considered more acceptable than swerving. In fact, even when the collision likelihood with the pedestrian was 80%, both actions were considered equally acceptable for AVs; only for human drivers swerving was considered slightly more acceptable.

Second, judgments of acceptability were quite different under uncertainty, where the likelihood of colliding with the bystander was unknown (Fig. 3, right). For AVs, the acceptability of both actions did not vary as a function of the pedestrian's collision likelihood; instead, staying in the lane was always considered more acceptable than swerving. A very similar pattern was obtained for human drivers, except when the pedestrian's collision likelihood was 80%; here, both actions were considered equally acceptable.

Taken together, these results indicate that the moral acceptability of actions can vary systematically with known likelihoods of a collision. Under conditions of uncertainty, however, staying in the lane seemed to be generally more acceptable than swerving, regardless of the stated likelihood of colliding with the pedestrian in the street.

3.3.3. Probability Estimates for Colliding with Bystander

In the uncertainty conditions, subjects were asked to estimate the likelihood of colliding with the bystander when swerving (on a scale of 0–100%, in steps of 1%). One possible outcome was that subjects would follow Laplace's principle of indifference and assign each outcome (collision vs. no collision) an equal probability (i.e., 50%) in the absence of distinguishing information. Another possibility was that estimates are influenced by the probability of colliding with the pedestrian. For instance, subjects may assume that the distance to the bystander is similar to the distance to the pedestrian and, therefore, assume a similar likelihood of being hit by the car. In this case, subjective estimates for the bystander should increase with the probability of colliding with the pedestrian.

The estimates (Fig. 4) in fact increased with the (known) likelihood of colliding with the pedestrian, for both the AV and human driver conditions. An ANOVA with driver type and likelihood of colliding with the pedestrian as between-subjects factors revealed a main effect of likelihood, $F(2, 417) = 31.79$, $p < 0.0001$, $\eta^2 = 0.13$; there was no influence of driver type and no significant interaction (both $p > 0.39$).

Overall, 36% (152 of 423) of subjects assigned a collision likelihood of exactly 50% to the bystander. Fig. 5 shows the decision preferences of this subgroup in comparison with the subjects in the risk condition, who were explicitly told that the likelihood of colliding with the bystander was 50%. Interestingly, subjects had the same decision preferences regardless of whether the likelihood was explicitly stated or subjectively assumed. Together, the results suggest that even under uncertainty, (subjective) probabilities may have entered subject's decision preferences. Consequently, 372 of 423 subjects (88%) stated that they considered the likelihood of colliding with the bystander when deciding how an AV or human driver should behave.

3.3.4. Decision Thresholds

We elicited the threshold probabilities at which subjects considered it acceptable to swerve. Subjects could either state that one should never swerve regardless of the likelihood of colliding with the pedestrian or provide a numerical estimate constituting the threshold value.

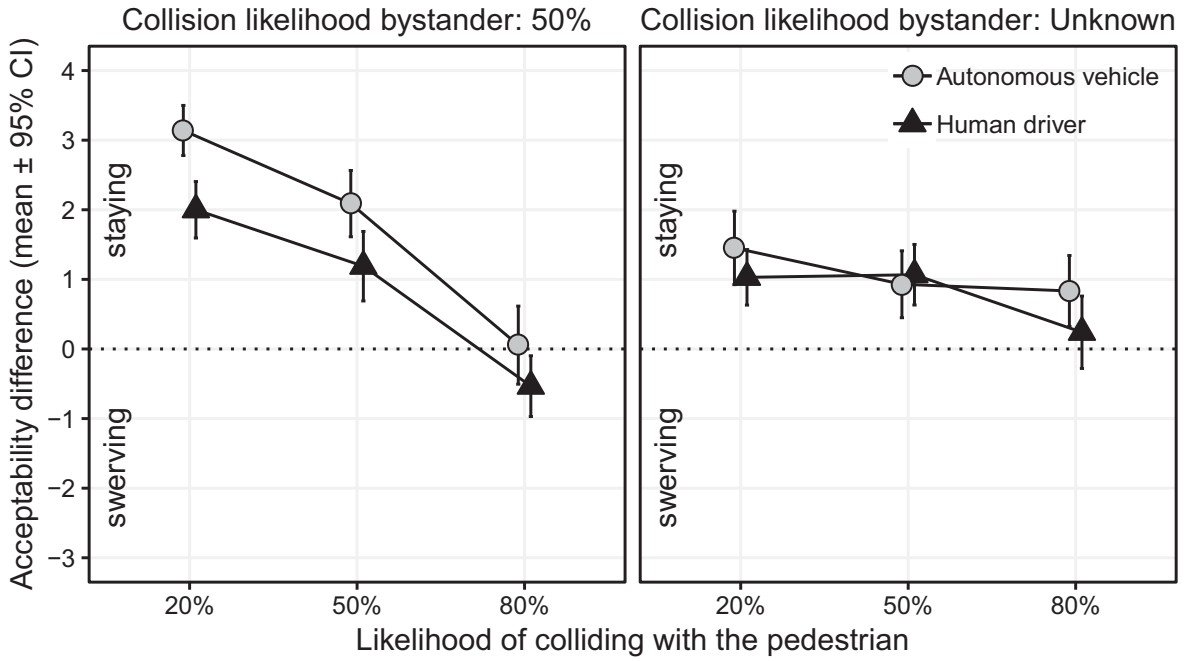

Fig. 3. Experiment 1. Difference in moral acceptability of staying in lane and swerving, under risk and uncertainty, as a function of likelihood of collision with the pedestrian in the street.

Table II. Results of an ANOVA for Moral Acceptability Judgments in Experiment 1; Dependent Variable Was the Within-Subject Difference Between the Moral Acceptability of Staying Minus Swerving

	df	SS	MS	*F*	*p*	η^2
Driver	1	74.06	74.06	18.44	<0.001	0.014
Likelihood pedestrian	2	464.89	232.44	57.87	<0.001	0.087
Likelihood bystander	1	34.32	34.32	8.54	0.004	0.006
Driver × Likelihood pedestrian	2	5.75	2.88	0.72	0.489	0.001
Driver × Likelihood bystander	1	18.43	18.43	4.59	0.032	0.003
Likelihood pedestrian × Likelihood bystander	2	167.49	83.75	20.85	<0.001	0.03
Driver × Likelihood pedestrian × Likelihood bystander	2	10.33	5.17	1.29	0.277	0.002
Error	860	3454.53	4.02			

Note: SS = type III sum of squares, MS = mean square.

Fig. 6(a) shows the proportion of subjects who responded that one should never swerve ($N = 221$). Under risk, this proportion was similar for human drivers and AVs, $\chi^2(1, N = 449) = 1.83$, $p = 0.18$ (aggregated across levels of the pedestrian's collision likelihood). Under uncertainty, more subjects responded that AVs should never swerve than that human drivers should never do so, $\chi^2(1, N = 423) = 9.64$, $p = 0.002$. Supporting our previous results, a substantial proportion of subjects preferred the stay option regardless of the likelihood of collision with the pedestrian, and under uncertainty this proportion was higher for AVs than for human drivers.

Fig. 6(b) shows the probability thresholds for people who stated that the car should swerve if a particular likelihood of collision with the pedestrian in the street was exceeded ($N = 651$). From a consequentialist perspective, one should swerve if it is higher than the likelihood of colliding with the bystander. Thus, in the risk conditions in which the likelihood of colliding with the bystander was known to be 50%, this threshold should be 51% (regardless of the stated risk for the pedestrian in the given situatio). However, only a minority of subjects gave exactly this threshold (38 of 346), and only 76 out of 346 subjects gave a threshold between 51% and 60% (inclusive). Instead, the estimates varied systematically

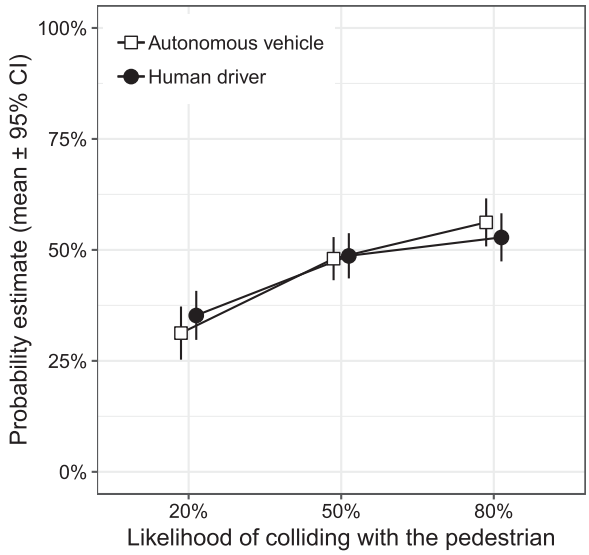

Fig. 4. Experiment 1. Subjects' estimates of likelihood of colliding with the bystander when swerving, under uncertainty.

with the risk for the pedestrian. An ANOVA with driver type and likelihood of collision with the pedestrian as between-subjects variables yielded a main effect of likelihood, $F(2, 340) = 9.53$, $p < 0.0001$, $\eta^2 = 0.05$, but no effect of driver type and no interaction.

Under uncertainty, the loss-minimizing threshold depends on the subjective probability of colliding with the bystander (Section 3.3.3). For each subject in the uncertainty conditions (excluding those who said that one should never swerve), we defined the loss-minimizing threshold as subjective P(bystander collision)+1%. Across all uncertainty conditions, only 2 of 305 subjects gave this answer; 40 out of 305 subjects gave a threshold estimate between 1% and 10% (inclusive) above P(bystander collision). Again, threshold judgments varied systematically as a function of risk for the pedestrian, $F(2, 299) = 22.3$, $p < 0.0001$, $\eta^2 = 0.13$; there was no effect of driver type and no interaction.

Note, however, that there was a high variance in the elicited threshold estimates across all conditions; we therefore advise caution in interpreting these data. One reason for the noisy estimates might be that subjects had difficulties in correctly understanding the question or possible consequences for the vehicle's behavior. Another possibility is that subjects can come up with some threshold estimate when directly being asked to do so, but that such estimates are not necessarily based on a consequentialist analysis of the situation. Rather, they may reflect other considerations and decision rules, which we discuss in the next section.

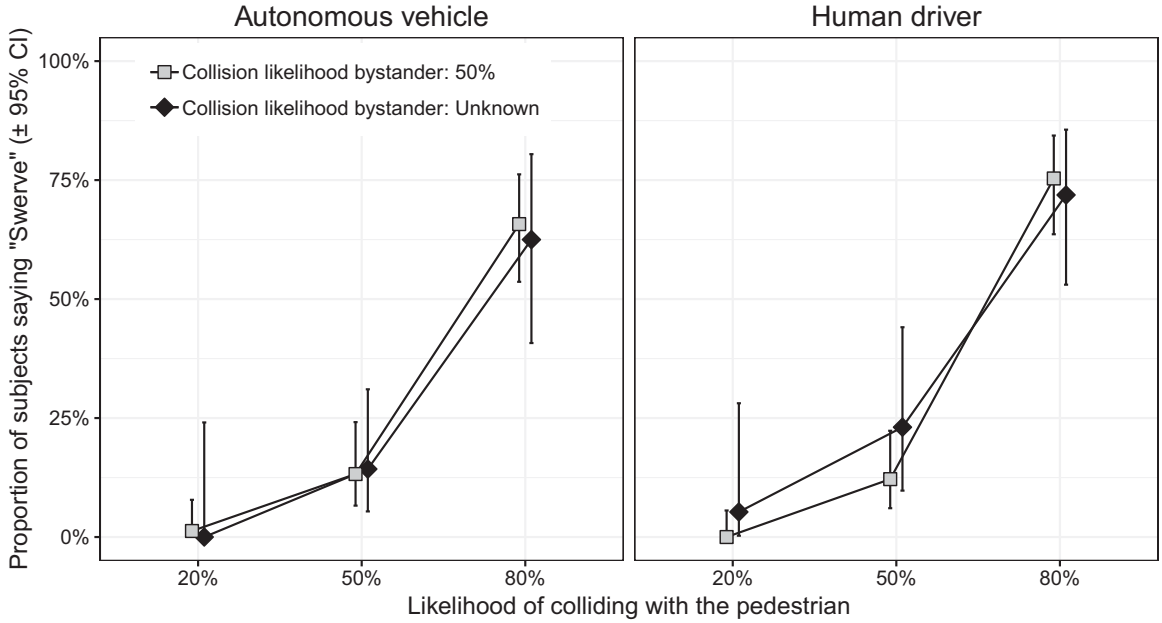

Fig. 5. Experiment 1. Decision preferences of subjects who assigned a subjective 50% likelihood to colliding with the bystander were virtually identical to decisions preferences in the risk condition, in which the likelihood of 50% was explicitly stated.

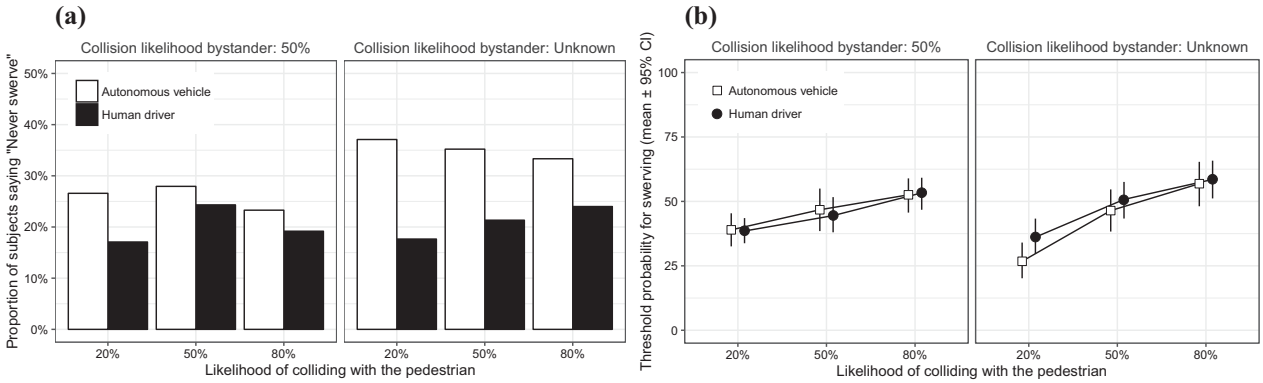

Fig. 6. Experiment 1. Swerving thresholds. (a) Proportion of subjects stating that one should never swerve, regardless of the likelihood of colliding with the pedestrian. (b) Mean estimates of subjects saying that one should only swerve if the likelihood of colliding with the pedestrian is higher than this particular probability.

3.3.5. Decision Rules

We asked subjects to "formulate a rule on how self-driving cars [drivers] should behave in situations like the one presented in this study." Their descriptions were analyzed as follows: first, author NK identified different kinds of rules inductively and coarsely classified these rules according to the cue(s) utilized, such as the likelihood of colliding with the pedestrian and bystander, respectively (Table III). Rules could use one of the cues (*single-cue rules*, SCs), more than one cue (*multiple-cue rules*, MCs), or no cues (*noncue rules*, NCs). We also included rules that were more idiosyncratic (*other rule*, O1–O3) or were not clearly identifiable (O4).

Second, two research assistants not involved in the project independently coded subjects' statements using this scheme, assigning each statement to one of the categories. Coders marked all cases for which they were uncertain. In a second round, both coders independently reexamined the critical cases that they did not classify in the first round. The interrater reliability between the two coders after this step was $\kappa = 0.65$. In a third round, the two coders jointly discussed cases with diverging classifications and cases where coders were still uncertain; they reached an agreement in 840 of 872 cases (interrater reliability $\kappa = 0.96$).

The two most frequent rules (Table III) were a consequentialist rule (MC1: "Always choose the option with the lowest (objective or subjective) likelihood of injuring somebody") and a simple default rule (NC1: "Always stay in the lane"). Another frequent rule was "Only swerve if the likelihood of hitting the bystander is 0%" (SC4).[5] The distribution of the two most frequent rules differed under risk and uncertainty. Under risk, the consequentialist rule was most frequent, whereas under uncertainty, it was mentioned substantially less often and tied with the default rule. A logistic regression focused on the two most frequent rules (MC1 and NC1) as dependent variable, considering only the subset of subjects who came up with either of these two rules (395 out of 840 subjects). The factors driver type and likelihood bystander served as predictors. This analysis showed that people were more likely to come up with a simple default rule under uncertainty than under risk, whereas driver type had no substantial influence (Table IV).

3.4. Summary and Conclusions

The key finding is that subjects placed particular weight on the default action (stay in the lane), even when it was not loss minimizing. This tendency was seen when the outcomes of the alternative action were uncertain, but also observed under risk when probabilities were known. Although probabilities may have entered subject's preferences, the stay option received more weight. The stay option was also considered morally more acceptable, especially under uncertainty and for AVs; its moral evaluation was stable across situations. The elicited decision

[5]The proportion of nonclassifiable rules (O4) ranged between 23% and 27%. Examples include such statements as "I don't trust self-driving cars" and "I believe self-driving cars should always take morally acceptable behavior and should follow the traffic guidelines."

Table III. Types of Decision Rules Provided by Subjects in Experiment 1; Percentages Indicate Rule Proportions Under Risk and Uncertainty, for Autonomous Vehicles (AV) and Human Drivers (HD); Percentages Based on the 840 out of 872 Cases for Which the Coders Agreed on the Classification

		Risk Bystander: 50%		Risk Bystander: Unknown	
Rule	Description	AV	HD	AV	HD
NC1	Always stay in the lane	18.7%	23.4%	19.5%	20.1%
NC2	Always swerve to the right	1.4%	2.7%	2.0%	3.8%
SC1	Only swerve if the likelihood of hitting the pedestrian is above a certain threshold/very high	2.4%	1.4%	5.0%	3.8%
SC2	Only swerve if the likelihood of hitting the bystander is below a certain threshold/very low	1.4%	2.7%	4.0%	1.0%
SC3	Only swerve if the likelihood of hitting the pedestrian is 100%	0.5%	1.8%	0.5%	1.0%
SC4	Only swerve if the likelihood of hitting the bystander is 0%	8.1%	4.5%	9.0%	9.1%
MC1	Always choose the option with the lowest (objective or subjective) likelihood of injuring somebody	34.4%	35.1%	14.0%	21.5%
MC2	Only swerve if the likelihood of hitting the bystander is substantially lower than the likelihood of hitting the pedestrian	3.8%	1.8%	1.0%	0.0%
MC3	Only swerve if the likelihood of hitting the pedestrian is above a certain threshold and if the likelihood of hitting the bystander is below a certain threshold	1.0%	1.4%	3.0%	1.0%
MC4	Only swerve if the likelihood of hitting the pedestrian is above a certain threshold/very high and if it is higher than the likelihood of hitting the bystander	1.0%	1.4%	2.0%	1.4%
MC5	Stay in the lane if the likelihood of injuring the bystander is unknown; if both likelihoods are known, choose the option with the lowest (objective or subjective) likelihood of injuring somebody	0.0%	0.0%	5.5%	4.3%
MC6	Swerve to the right if the likelihood of injuring the bystander is unknown; if both likelihoods are known, choose the option with the lowest (objective or subjective) likelihood of injuring somebody	0.0%	0.0%	1.5%	1.0%
O1	The driver should take control of the car	0.5%	0.0%	0.0%	0.0%
O2	The driver should follow his or her gut feeling/instincts/intuitions	0.0%	0.5%	0.0%	2.9%
O3	Alternative rule developed by the subject	1.9%	0.9%	6.5%	1.9%
O4	Statement does not contain a clearly defined decision rule	24.9%	22.5%	26.5%	27.3%

Note: Rules were coarsely classified according to the cues (pieces of information) they utilize, such as the likelihood of colliding with the pedestrian and bystander, respectively. NC = noncue rule; SC = single-cue rule; MC = multiple-cue rule; O = other rule.

rules further support that subjects preferred the stay option as a default, especially under uncertainty.

The tendency toward a default of staying is consistent with decision-making research in other domains (Benartzi & Thaler, 2013; Johnson & Goldstein, 2003; Madrian & Shea, 2001; Pichert & Katsikopoulos, 2008; Sunstein, 2017). Although following the default may lead to suboptimal behavior in particular cases, there are several rationales in its favor in the case of driving. First, it is certainly safer to stick to the general rules of driving and traffic regulations. Deviating from norms may lead to a better outcome in a specific case, but regularly doing so makes traffic unpredictable and uncertain. Second, some actions, such as staying in the lane and making an emergency stop, are easier to implement and control and

Table IV. Results of a Bayesian Logistic Regression with the Two Most Frequent Rules as Dependent Variable (Coded 0 = MC1 = "Always Choose the Option with the Lowest [Objective or Subjective] Likelihood of Injuring Somebody," 1 = NC1 = "Always Stay in the Lane")

	Coefficient	*SE*	*p*
Intercept	–0.594	0.195	0.002
Driver	0.182	0.261	0.486
Likelihood bystander	0.900	0.308	0.003
Driver × Likelihood bystander	–0.549	0.407	0.177
N			395
McFadden's R^2			0.02

Note: Dummy variable for driver, coded 0 = autonomous vehicle and 1 = human driver. Dummy variable for likelihood bystander, coded 0 = 50% and 1 = unknown. Estimates based on Bayesian logistic regression using independent Cauchy priors with center 0 and scale 2.5 (Gelman & Su, 2016; Gelman et al., 2008).

thus less uncertain in their consequences. Third, a simple default not only requires less information and offers a higher degree of control, but behavior is also more easily and transparently evaluated after the decision. The retrospective evaluation of behavior is an important issue, as accidents with AVs are bound to happen and will inevitably lead to investigations and public discussion of appropriate policy standards for AVs.

4. EXPERIMENT 2

From a policy perspective, it is desirable that AVs implement actions that are societally acceptable even if accidents occur. Experiment 2 tested the hypothesis that the evaluation of the default stay option is stable even when the AV has actually harmed another road user.

From a normative perspective, retrospective evaluations are often problematic because outcome information is known to affect people's retrospective judgments—even if it was not available at the time of the decision and should therefore be ignored. Studies on outcome bias show that actions are judged worse when they led to bad rather than good outcomes in an otherwise identical situation (Baron & Hershey, 1988; Sezer, Zhang, Gino, & Bazerman, 2016). Studies on hindsight effects have demonstrated how judged probabilities are affected: once people know the outcome in hindsight, they tend to believe that an event was more predictable than they would have thought beforehand (Fischhoff, 1975; Hawkins & Hastie, 1990). Observing a specific outcome may thus indirectly change the decision evaluation by increasing the perceived probability of this outcome or by directly influencing the evaluation of the decision. Hindsight and outcome effects have already been observed in moral dilemmas under uncertainty (Fleischhut et al., 2017).

However, some actions may be more robust against outcome bias and hindsight effects than others. For instance, a consequentialist rule considers the likelihood of the consequences yet may lead to a different evaluation if probabilities or outcomes are judged differently in hindsight. In contrast, a simple default rule that does not consider alternative consequences or probabilities may be evaluated similarly before and after the fact. This suggests that the moral evaluation of staying in the lane and initiating an emergency stop as a default may be less susceptible to outcome effects (i.e., collision with another road user) than a swerving maneuver. We tested this hypothesis in situations of risk, in which the AV could estimate collision likelihoods, and in situations of uncertainty, in which the AV could not. Thus, whereas in Experiment 1 the uncertainty manipulation only pertained to the bystander (i.e., the likelihood of colliding with the bystander was known or unknown), in Experiment 2 we also investigated scenarios in which the likelihood of colliding with the pedestrian on the street was unknown. In the uncertainty conditions, we additionally tested whether outcome information influenced subjects' probability estimates in hindsight.

4.1. Subjects and Design

Subjects were recruited via the AMT platform and paid $0.80 USD for a mean of 8.2 minutes. We varied three factors: type of decision situation (risk vs. uncertainty), which action was taken by the AV (stay vs. swerve), and resulting outcome (collision vs. no collision), yielding a 2 × 2 × 2 between-subjects design.

4.2. Procedure and Materials

The scenario was largely identical to that in Experiment 1: a dilemma situation in which staying in the lane endangers a pedestrian in the street and swerving to the right endangers a bystander on the sidewalk (Fig. 7). In the *risk conditions*, the collision likelihood for both the pedestrian and the bystander was 50%, as estimated by the car's systems. In the *uncertainty conditions*, no likelihood information

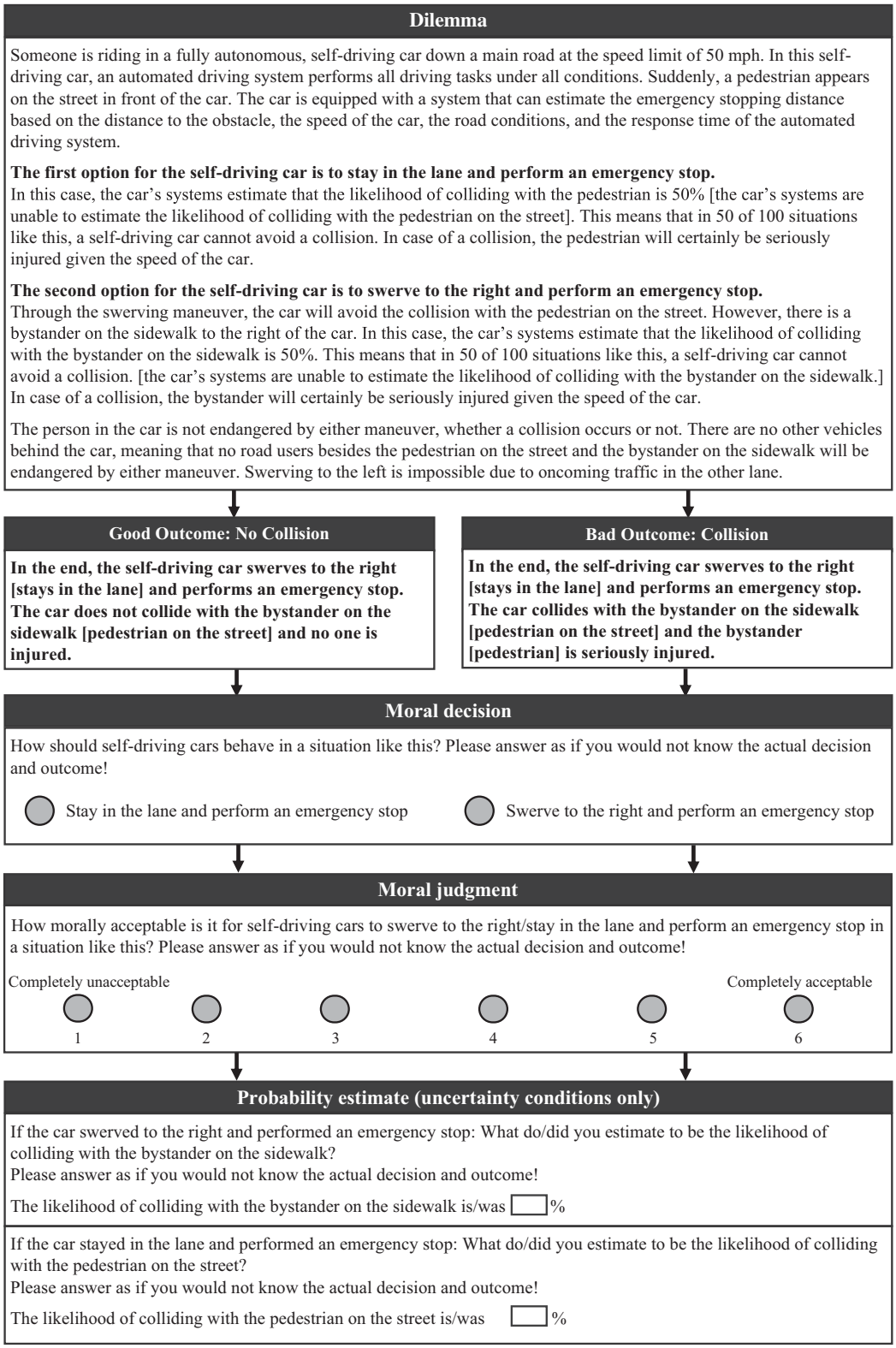

Fig. 7. Overview of experimental procedure and information given to subjects in Experiment 2. Words in brackets indicate variations between conditions; order of answer options was randomized across subjects.

was provided because the car's systems were unable to estimate collision likelihoods for either the pedestrian or the bystander. Unlike in Experiment 1, we informed subjects about the action taken by the AV and the consequence that occurred (collision vs. no collision). Subjects who twice failed the initial instruction test were not allowed to participate.

Subjects answered three questions. The first asked for a *decision preference*: "How should self-driving cars behave in a situation like this?" Possible answers were staying in the lane and swerving to the right. The second question requested a *moral judgment*, asking subjects to evaluate the moral acceptability of both actions on a scale of 1 (*completely unacceptable*) to 6 (*completely acceptable*). In the uncertainty conditions, in which the collision likelihoods were not specified (Fig. 7), we additionally asked subjects to estimate the collision likelihood for the pedestrian and the bystander. For each question subjects were instructed: "Please answer as if you would not know the actual decision and outcome!"

4.3. Results

The task was accepted by 1,761 subjects on the AMT platform, who were randomly assigned to the conditions. Of these, 621 failed the instruction test and 366 dropped out during the survey. Three subjects were excluded because they indicated they had participated before, and five because of a self-reported limited understanding of English. The analyses are therefore based on $N = 766$ subjects; with between 91 and 100 subjects in the different conditions.

4.3.1. How Should an AV Behave in a Situation Like This?

Fig. 8 shows the proportion of people who preferred the AV to swerve, as a function of the action that was taken (stay or swerve) and its consequence (no collision vs. collision). Table V summarizes the results of a logistic regression.

The key finding is an asymmetry between preferences for the stay and the swerve action under risk and uncertainty, as indicated by the significant Action × Outcome interaction. If the car stayed in the lane, the outcome did not affect subjects' judgments of how an AV should behave in such a situation. Irrespective of whether a collision occurred or not, no more than 10% of subjects considered swerving to be the desirable action—under both risk, $\chi^2(1, N = 189) = 0.56$, $p = 0.45$, and uncertainty, $\chi^2(1, N = 191) = 1.76$, $p = 0.18$. By contrast, if the car swerved, the outcome strongly influenced subjects' preferences. If no collision occurred, about 40% preferred the AV to swerve, but less than 20% did when a collision occurred. This pattern was obtained under risk, $\chi^2(1, N = 191) = 9.09$, $p = 0.003$, and uncertainty, $\chi^2(1, N = 195) = 6.25$, $p = 0.01$.

Table V. Results of a Bayesian Logistic Regression for Decision Preferences in Experiment 2 (Coded 0 = Stay, 1 = Swerve)

	Coefficient	*SE*	*p*
Intercept	–2.478	0.353	<0.001
Situation	0.283	0.451	0.53
Action	0.997	0.419	0.02
Outcome	–0.580	0.527	0.27
Situation × Action	–0.304	0.696	0.66
Situation × Outcome	–0.247	0.546	0.65
Action × Outcome	1.615	0.593	<0.01
Situation × Action × Outcome	0.137	0.781	0.86
N			766
McFadden's R^2			0.12

Note: Dummy variable for situation, coded 0 = risk and 1 = uncertainty. Dummy variable for taken action, coded 0 = stay and 1 = swerve. Dummy variable for outcome, coded 0 = no collision and 1 = collision. Estimates based on Bayesian logistic regression using independent Cauchy priors with center 0 and scale 2.5 (Gelman & Su, 2016; Gelman et al., 2008).

The upshot is that the preferences for both actions were differentially affected by the outcome: whereas the decision outcome was largely irrelevant when the car stayed in the lane, it strongly influenced preferences about the swerve action. Thus, when accidents occurred, preferences for the default stay in lane option were more robust than preferences for the swerve option.

4.3.2. How Morally Acceptable Are the Two Actions?

As in Experiment 1, we computed for each subject the difference between the moral acceptability of staying and swerving, such that a positive difference indicates that staying in the lane was considered more acceptable and a negative difference indicates that swerving was considered more acceptable. Fig. 9 shows the mean differences as a function of the action that was taken (stay vs. swerve) and its consequences (no collision vs. collision); Table VI shows the results of an ANOVA with decision situation, taken action, and outcome as between-subject factors.

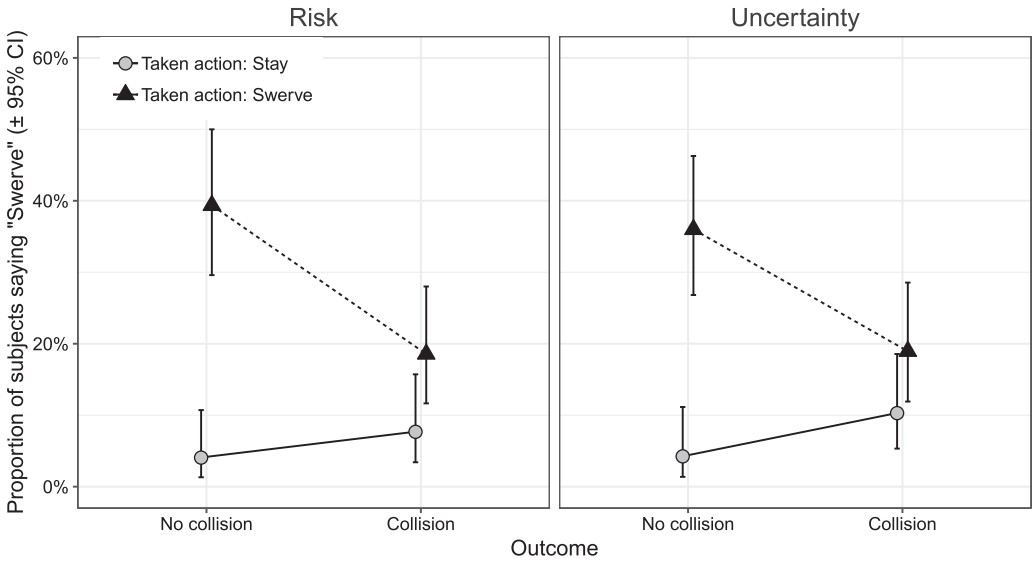

Fig. 8. Experiment 2. Decision preferences regarding how an AV should behave in the dilemma situation as a function of taken action (stay vs. swerve) and decision outcome (collision vs. no collision) for risk and uncertainty conditions.

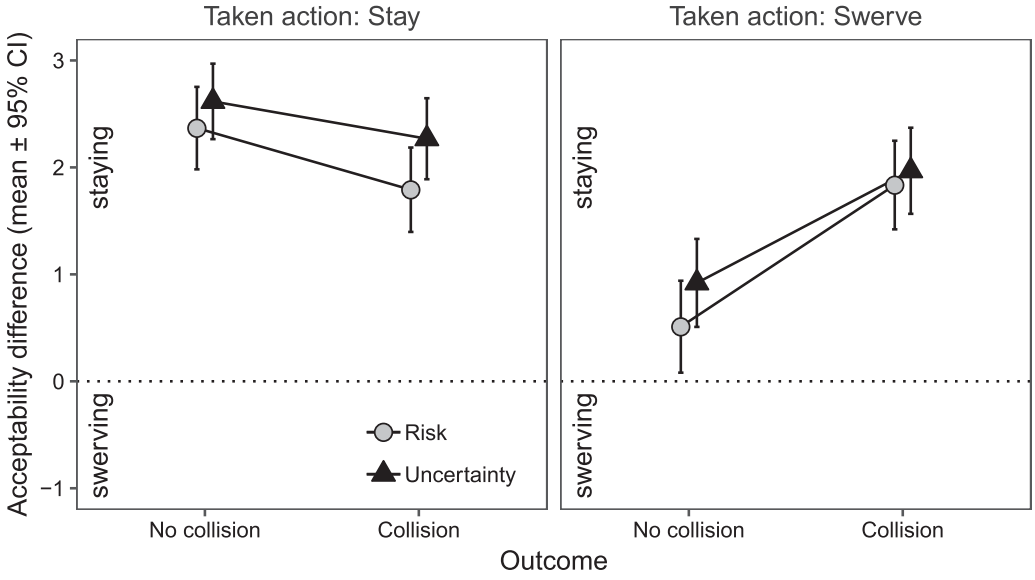

Fig. 9. Experiment 2. Difference in moral acceptability judgments as a function of situation (risk vs. uncertainty), taken action (stay vs. swerve), and decision outcome (no collision vs. collision).

Of particular interest is the extent to which the outcome (collision or no collision) influenced the moral acceptability of the actions (Fig. 9). If the car stayed in the lane, a collision had very little influence on judgments of moral acceptability, with people generally considering staying in the lane as more acceptable than swerving. Thus, even when the taken action led to a collision, staying in the lane was considered more acceptable than swerving; this was observed both under risk and uncertainty. By contrast, if the car had swerved, the resulting outcome strongly influenced the relative moral acceptability of the two actions. If no collision occurred, people judged the stay action to be slightly more acceptable than the swerve action. However, if a collision did occur as a result of the swerving maneuver, staying in the lane was considered much more morally acceptable than swerving, both under risk and uncertainty.

Table VI. Results of an ANOVA for Moral Acceptability Judgments in Experiment 2; Dependent Variable Was the Within-Subject Difference Between the Moral Acceptability of Staying Minus Swerving

	df	SS	MS	*F*	*p*	η^2
Situation	1	19.27	19.27	5.03	0.025	0.003
Action	1	173.56	173.56	45.30	<0.001	0.03
Outcome	1	25.07	25.07	6.54	0.011	0.004
Situation × Action	1	0.40	0.40	0.11	0.746	0.000
Situation × Outcome	1	0.03	0.03	0.01	0.931	0.000
Action × Outcome	1	130.08	130.08	33.95	<0.001	0.02
Situation × Action × Outcome	1	3.03	3.03	0.79	0.374	0.001
Error	758	2904.17	3.83			

Note: SS = type III sum of squares, MS = mean square.

4.3.3. Probability Estimates Under Uncertainty

Subjects in the uncertainty conditions provided estimates for the likelihood of colliding with the pedestrian and the bystander. Although such estimates are admittedly difficult in the absence of any further information in the scenario description, they allow testing whether people show a hindsight effect, that is, adjust their subjective probability estimates toward the outcome even when asked to ignore this information. A hindsight effect would predict that people estimate the collision with the pedestrian or bystander to be more likely if it in fact occurred than if it did not occur. It is less clear whether one should also expect a hindsight effect for the road user not put at risk by the action (i.e., the bystander if the car stays in the lane, and the pedestrian if the car swerves). These estimates might be unaffected by the collision or noncollision with the person at risk, or people might take the outcome as a cue for their collision estimates for the person currently not at risk (e.g., by assuming a similar distance to the approaching car).

The results show a hindsight effect for the collision likelihood for the bystander as well as for the pedestrian, irrespective of the action taken and the person at risk (Fig. 10). The outcome influenced the estimated likelihood for the pedestrian when the car stayed in the lane, $t(189) = 6.39$, $p < 0.0001$, $d = 0.92$, as well as when the car swerved, $t(193) = 4.71$, $p < 0.0001$, $d = 0.67$. Similarly, the outcome influenced subjects' estimates for the bystander when the car stayed in the lane, $t(189) = 3.27$, $p = 0.001$, $d = 0.47$, as well as when the car swerved, $t(193) = 6.5$, $p < 0.0001$, $d = 0.93$. Thus, subjects' probability estimates in hindsight varied depending on how events unfolded.

4.4. Summary and Conclusions

The key finding of Experiment 2 is that the evaluation of the default stay in lane option was robust against hindsight and outcome effects when accidents occurred. When the car stayed in the lane, decision preferences and moral judgments did not substantially change regardless of whether a collision occurred or not, whereas a collision affected both when resulting from a swerve maneuver.

One explanation for this pattern may be that the simple default requires no information about alternative actions or probabilities; therefore, its evaluation is less likely to change even when outcomes are known in hindsight. A different psychological mechanism underlying this finding may be grounded in the asymmetry of counterfactuals. Research has shown that people tend to consider counterfactual alternatives for norm-*violating* behavior but not for norm-*conforming* behavior (Petrocelli, Percy, Sherman, & Tormala, 2011; Philips, Luguri, & Knobe, 2015). In the case of driving, people may thus not consider counterfactual outcomes when the AV conforms to traffic norms (i.e., stays in the lane as a default), but may activate counterfactual possibilities when the AV transgresses these norms (i.e., swerves). Preferentially considering norm-conforming behavior can thus lead to an increased weight of the default course of actions.

5. GENERAL DISCUSSION

We have argued that dilemma scenarios used to investigate moral judgments about the behavior of AVs should reflect the conditions under which AVs operate in real-world environments. This is important to ensure a match between the situations that are used to investigate what is acceptable to the public and the characteristics of the actual decision problem. Our results show that people respond differently depending on the degrees of risk and uncertainty inherent in the situation, and that their preferences for different types of decision rules vary accordingly. The first key finding of the studies is a general preference for staying in the lane. This result supports the idea that people consider the stay option to be a reasonable default in critical traffic situations, even if it does not minimize expected losses. Although the default action may be inconsistent with a

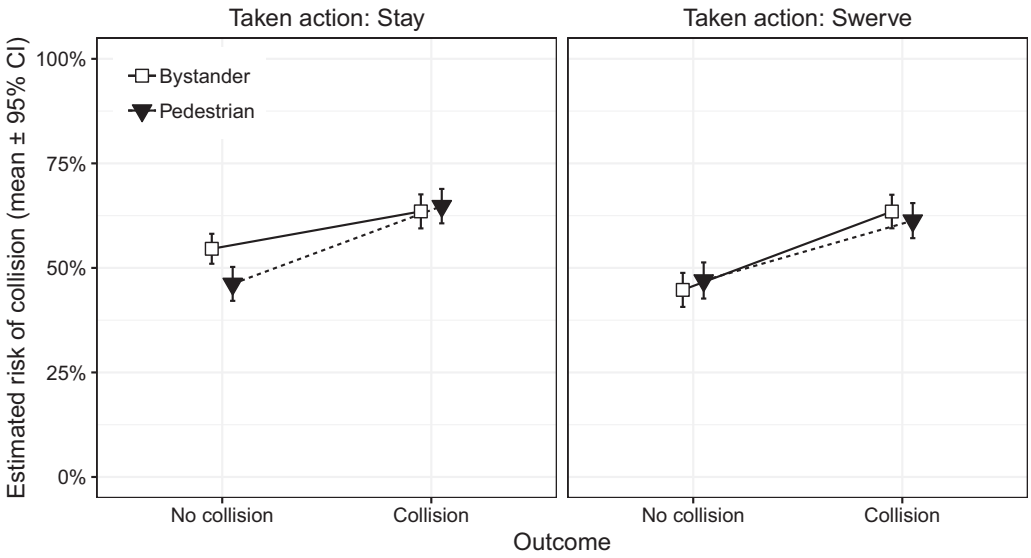

Fig. 10. Experiment 2. Estimated collision likelihoods for pedestrian and bystander under uncertainty.

simple consequentialist account in specific situations, it is defensible from other perspectives. According to consequentialism, the morally right act is the one that maximizes some social utility criterion, where the utility criterion is often applied directly to a single action (e.g., expected number of people saved given each action). Yet from a *rule-consequentialist perspective*, the rightness of an action derives not from its consequences in a single case, but from whether the action maximizes utility in a class of situations governed by the rule (Hooker, 1990). Typically, these rules are assumed to be fairly general, representing societally accepted values (e.g., not torturing or murdering). An individual action can be morally right even when suboptimal in a specific situation, as long as it complies with a rule that is assumed to maximize social welfare if followed by everybody. Moreover, a simple default rule often leads to better results as it requires less information and provides a higher degree of control. In the present studies, traffic norms set defaults that may serve as the relevant rules in the moral evaluation of actions. When children learn how to navigate traffic, they are taught to use the sidewalk to stay safe, based on the rule that vehicles have to stay on the road. When drivers learn how to react to sudden obstacles in the street, they are taught to stay in their lane and brake. Transgressing the default may lead to better outcomes in particular situations but as a general policy would be suboptimal (Bicchieri, 2005) and could undermine road users' coordination, which might easily lead to an increase in accidents.

The second key finding is that the evaluation of the default action was more robust against hindsight and outcome effects than the swerving action when an accident occurred: a bad outcome had less influence on the moral evaluation of conforming to the default norm (i.e., staying in the lane) than on the evaluation of deviating from the norm (i.e., swerving). Often, outcomes enter moral evaluations even when they are due to probabilistic factors outside the control of the agents (moral luck) (Cushman, 2008; Young, Nichols, & Saxe, 2010). The present experiments extend these findings by demonstrating an asymmetry in the evaluation of norm-deviating and norm-conforming (default) behavior. Outcome and hindsight biases influence moral evaluations more strongly when the target behavior deviates from the norm than when it conforms to the norm.

5.1. Implications

Our results highlight the importance of investigating moral judgment and decision making in situations in which consequences are only probabilistically known or not precisely quantified. The case of AVs provides an interesting area of inquiry because it relates theoretical debates about appropriate ethical standards to applied decision-making research. Recently, there has been renewed interest

in the psychology of moral judgment across different fields, including economics, the cognitive and decision sciences, neuroscience, and experimental philosophy (Bartels & Medin, 2007; Bennis, Medin, & Bartels, 2010; Fehr & Gächter, 2000; Gigerenzer, 2010; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Tan, Luan, & Katsikopoulos, 2017; Wiegmann & Waldmann, 2014). Studies on the ethics of AVs should not be limited to dilemma situations with fixed outcomes and a restricted set of judgments but should harness the rich literature that has emerged across disciplinary boundaries, for instance, to empirically investigate retrospective judgments of blame and responsibility (Gerstenberg & Lagnado, 2014; Lagnado & Channon, 2008) for AV-related accidents.

Our research also highlights practical key issues for the kinds of AV decision policies that are desirable, implementable, and societally acceptable. Oftentimes, some kind of consequentialist account is endorsed as the normative yardstick. Yet the first challenge to such an approach arises from information constraints: What if the required information for a consequentialist analysis is not available? Note that this issue is relevant for other approaches, too. Consider a decision rule such as "Only swerve if the likelihood of hitting the bystander is 0%," which was frequently mentioned in Experiment 1. Following this rule also requires information about possible decision consequences, namely, certainty that the action does not put a noninvolved person at risk. Taking seriously the distinction between different types and degrees of risk and uncertainty helps to better characterize the information requirements of different decision policies and the circumstances under which particular accident algorithms might perform well or poorly.

A second challenge arises from valuing decision outcomes in a systematic and societally acceptable fashion. For instance, the German Ethics Commission on Automated Driving (2017) discussed dilemma situations in its report and concluded:

> In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. It is also prohibited to offset victims against one another. General programming to reduce the number of personal injuries may be justifiable. Those parties involved in the generation of mobility risks must not sacrifice noninvolved parties. (p. 11)

Thus, broader legal and ethical considerations also put constraints on possible decision policies.

A third challenge lies in how a person's willingness to adopt new technologies and policies depends on how well they align with that person's self-interest. For instance, in one study (Bonnefon et al., 2016), subjects endorsed a consequentialist viewpoint, showing a preference for AVs that would sacrifice their own passengers to minimize the total number of casualties. But they wanted their own AVs to put a premium on passenger safety, rather than maximizing the greater good. This tension between the general moral principles that people endorse and personal self-interest highlights the need to find policies that are societally acceptable even under critical conditions.

5.2. Limitations and Future Research

The present studies represent a first step toward investigating judgment and decision making about AVs under risk and uncertainty. One key goal for future research is to disentangle possible explanations for the default effect. For instance, people may assign different levels of blame to the involved parties (e.g., because they may assume that the pedestrian in the street has endangered himself or herself by stepping into the street), focus on the actions' different degrees of control, or generally refuse to pit victims against each other.

Another issue for future research is to explore the relationships between moral preferences (or decisions) and moral judgments. In our studies, we first asked subjects to indicate their moral preference (i.e., whether the AV or human driver should stay in the lane or swerve) because this was our key variable of interest. Expressing such a preference may have an impact on subsequent judgments though, such as the moral acceptability of the different options. Future studies should investigate this issue in more detail, in order to advance our understanding of the relationships between judgment and decision making in the moral domain.

A key question concerns the interplay between behavioral studies and policy making. We used a reputation-based criterion (95% approval rate on the MTurk platform, which has been shown to lead to higher quality data than attention-check questions) (Peer et al., 2014), as well an instruction test to make sure that all subjects carefully read the given scenario. This ensures internal validity, but for policy making external validity is just as important. Although online studies tend to have more diverse samples than standard lab studies (e.g., in terms of

age, ethnicity, and education) (Berinsky et al., 2012; Casler et al., 2013), they are often not representative (but see Levay et al., 2016). Ultimately, if behavioral research is supposed to guide policy making, representative studies should be conducted. For instance, while our analyses of the attrition rates indicate that there was no imbalance with respect to the demographic factors we elicited, we cannot rule out that the drop-out rates were affected by unobserved variables. Conducting studies with representative samples would address this issue.

Another strategy to achieve higher external validity is to use large-scale crowd-sourcing platforms, such as the Moral Machines website (https://moralmachine.mit.edu; cf. Bonnefon et al., 2016). This may reduce internal validity but allows collecting data on a broad array of scenarios and a potentially diverse set of users. It can also foster public debate, as it allows users to design their own scenarios and discuss ideas and judgments with other people. It would be straightforward to amend existing scenarios (which typically present all decision consequences as certain) with varying degrees and types of uncertainty to achieve external validity with respect to not only the sampled population, but also the circumstances under which AVs have to operate in the real world.

ACKNOWLEDGMENTS

This research was partially supported by grant ME 3717/2-2 to BM from the Deutsche Forschungsgemeinschaft as part of the priority program "New Frameworks of Rationality" (SPP 1516). We thank Charley M. Wu for helpful comments, Clara Schirren and Clara Brune for help with coding the decision rules in Experiment 1, and Anita Todd and Susannah Goss for editing the article. Data and R code available at https://osf.io/sx85t/.

REFERENCES

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–579.

Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, *18*, 24–28.

Benartzi, S., & Thaler, R. H. (2013). Behavioral economics and the retirement savings crisis. *Science*, *339*, 1152–1153.

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives in Psychological Science*, *5*, 187–202.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical Turk. *Political Analysis*, *20*, 351–368.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*, 1573–1576.

Boudette, N. E. (2017). January 19. Tesla's self-driving system cleared in deadly crash. *New York Times*. Retrieved from http://www.nytimes.com/2017/01/19/business/tesla-models-autopilot-fatal-crash.html.

Brand, C. M., & Oaksford, M. (2015). The effect of probability anchors on moral decision making. In D. C. Noelle, R. Dale, A. S. Warlaumon, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 268–272). Austin, TX: Cognitive Science Society.

Camerer, C., & Weber, W. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*, 325–370.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*, 2156–2160.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, *75*, 643–669.

Ethics Commission Automated and Connected Driving (2017). Federal Ministry of Transport and Digital Infrastructure. Retrieved from https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*, 980–994.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 288–299.

Fleischhut, N., Meder, B., & Gigerenzer, G. (2017). Moral hindsight. *Experimental Psychology*, *64*, 110–123.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.

Fraichard, T., & Asama, H. (2004). Inevitable collision states—A step towards safer robots? *Advanced Robotics*, *18*, 1001–1024.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. -S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*, 1360–1383.

Gelman, A., & Su, Y-S. (2016). arm: Data analysis using regression and multilevel/hierarchical models. *R Package Version*, *1*, 9–3. Retrieved from https://CRAN.R-project.org/package=arm.

Gerdes, J. C., & Thornton, S. M. (2015). Implementable ethics for autonomous vehicles. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous driving* (pp. 87–102). Heidelberg: Springer.

Gerstenberg, T., & Lagnado, D. A. (2014). Attributing responsibility: Actual and counterfactual worlds. In J. Knobe, T. Lombrozo, & S. Nichols (Eds.), *Oxford studies of experimental philosophy* (Vol. *1*, pp. 91–130). Oxford: Oxford University Press.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, *2*, 528–554.

Goodall, N. J. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record*, *2424*, 58–65.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, *107*, 311–327.

Hern, A. (2016). August 22. Self-driving cars don't care about your moral dilemmas. *Guardian*. Retrieved from http://www.theguardian.com/technology/2016/aug/22/self-driving-cars-moral-dilemmas.

Hooker, B. (1990). Rule-consequentialism. *Mind*, *99*, 67–77.

Johnson, E. J., & Goldstein, D. G. (2003). Do defaults save lives? *Science*, *302*, 1338–1339.

Knight, F. H. (1921). *Risk, uncertainty and profit*. Boston: Houghton Mifflin.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*, 754–770.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *SAGE Open*, *6*, 1–17.

Lin, P. (2015). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous driving* (pp. 69–85). Heidelberg: Springer.

Madrian, B. C., & Shea, D. (2001). The power of suggestion: Inertia in 401(k) participation and savings behavior. *Quarterly Journal of Economics*, *116*, 1149–1187.

Malle, B. F., Scheutz, M., Arnold, T. M., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). ACM.

Meder, B., Le Lec, F., & Osman, M. (2013). Decision making in uncertain times: What can cognitive and decision sciences say about or learn from economic crises? *Trends in Cognitive Science*, *17*, 257–260.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, *19*, 1275–1289.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023–1031.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology*, *100*, 30–46.

Philips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Pichert, D., & Katsikopoulos, K. V. (2008). Green defaults: Information presentation and pro-environmental behaviour. *Journal of Environmental Psychology*, *28*, 63–73.

Powell, D., Cheng, P. W., & Waldmann, M. R.(2016). How should autonomous vehicles behave in moral dilemmas? Human judgments reflect abstract moral principles. In A. Papafragou, D. Grodner, D. Mirman D, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 307–312). Austin, TX: Cognitive Science Society.

Sezer, O., Zhang, T., Gino, F., & Bazerman, M. (2016). Overcoming the outcome bias: Making intentions matter. *Organizational Behavior and Human Decision Processes*, *137*, 13–26.

Shenhav, A., & Greene, J. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*, 667–677.

Singh, S. (2015). *Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey*. Traffic Safety Facts Crash Stats. Washington, DC: National Highway Traffic Safety Administration (US). Report No. DOT HS 812 115.

Sunstein, C. R. (2017). Default rules are better than active choosing (often). *Trends in Cognitive Science*, *21*, 600–606.

Tan, J. H., Luan, S., & Katsikopoulos, K. (2017). A signal-detection approach to modeling forgiveness decisions. *Evolution and Human Behavior*, *38*, 27–38.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. (2016). Moral judgments of human vs. robot agents. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 775–780). New York: IEEE.

Wakker, P. P. (2010). *Prospect theory for risk and ambiguity*. Cambridge: Cambridge University Press.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). Oxford: Oxford University Press.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*, 28–43.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*, 333–349.