

# Human Bias Towards Autonomous Vehicles

# Psychological roadblocks to the adoption of self-driving vehicles

Self-driving cars offer a bright future, but only if the public can overcome the psychological challenges that stand in the way of widespread adoption. We discuss three: ethical dilemmas, overreactions to accidents, and the opacity of the cars' decision-making algorithms — and propose steps towards addressing them.

Azim Shariff, Jean-François Bonnefon and Iyad Rahwan

The widespread adoption of autonomous vehicles promises to make us happier, safer and more efficient. Manufacturers are speeding past the remaining technical challenges to the cars' readiness. But the biggest roadblocks standing in the path of mass adoption may be psychological, not technological; 78% of Americans report fearing riding in an autonomous vehicle, with only 19% indicating that they would trust the car<sup>1</sup>.

Trust — the comfort in making oneself vulnerable to another entity in the pursuit of some benefit — has long been recognized as critical to the adoption of automation, and becomes even more important as both the complexity of automation and the vulnerability of the users increase<sup>2</sup>. For autonomous vehicles, which will need to navigate our complex urban environment with the power of life and death, trust will determine how widely the cars are adopted by consumers, and how tolerated they are by everyone else. Achieving the bright future promised by autonomous vehicles will require overcoming the psychological barriers to trust. Here we diagnose three factors underlying this resistance and offer a plan of action (see Table 1).

**The dilemmas of autonomous ethics**  
The necessity for autonomous vehicles to make ethical decisions leads to a series of dilemmas for their designers, regulators and the public at large<sup>3</sup>. These begin with the need for an autonomous vehicle to decide how it will operate in situations where its actions could decrease the risk of harming its own passengers by increasing the risk to a potentially larger number of non-passengers (for example, pedestrians and other drivers).

**Table 1 | A summary of the psychological challenges to autonomous vehicles, and suggested actions for overcoming them**

Psychological challenge	Suggested actions
<b>The dilemmas of autonomous ethics.</b> People are torn between how they want autonomous vehicles to ethically behave; they morally believe the vehicles should operate under utilitarian principles, but prefer to buy vehicles that prioritize their own lives as passengers. The idea of a car sacrificing its passengers deters people from purchasing an autonomous vehicle.	Shift the discussion from the relative risk of injury to the absolute risk. Appeal to consumers' desire for virtue signalling.
<b>Risk heuristics and algorithmic aversion.</b> The novelty and nature of autonomous vehicles will result in outsized reactions in the face of inevitable accidents. Such overreactions risk slowing or stalling the adoption of autonomous vehicles.	Prepare the public for the inevitability of accidents. Openly communicate algorithmic improvement. Manage public overreaction with 'fear placebos' and information about actual risk levels.
<b>Asymmetric information and the theory of the machine mind.</b> A lack of transparency into the underlying decision-making processes can make it difficult for people to predict the autonomous vehicles' behaviour, diminishing trust.	Research the type of information required to form trustable mental models of autonomous vehicles.

to spare the lives of two or more pedestrians, or vice versa (Fig. 1).

In handling these situations, the cars may operate as utilitarians, minimizing total risk to people regardless of who they are, or as self-protective, placing extra weight on the safety of their own passengers. Human drivers make such decisions instinctively in a split second, and thus cannot be expected to abide by whatever ethical principle they formulated in the comfort of their armchair. But autonomous vehicle manufacturers have the luxury of moral deliberation — and thus the responsibility of that deliberation.

The existence of this ethical dilemma in

the cars to save the greater number. But as consumers, they want self-protective cars<sup>4</sup>. As a result, adopting either strategy brings its own risks for manufacturers — a self-protective strategy risks public outrage, whereas a utilitarian strategy may scare consumers away.

Both the ethical and social dilemmas will need to be addressed to earn the trust of the public. And because it seems unlikely that regulators will adopt the strictest self-protective solution — in which autonomous vehicles would never harm their passengers, however small the danger to passengers and large the risk to others — we will have to

## Psychological roadblocks to the adoption of self-driving vehicles

The dilemma of autonomous ethics

Overreactions to inevitable AV accidents

Lack of transparency in the machine decision making process

# Roadblock 1: Dilemma of autonomous ethics

## Problem

- Utilitarian vehicles vs. self-protective vehicles

## Suggested Solutions

- Shift the discussion from the relative risk of injury to the absolute risk
- Appeal to consumers' desire for virtue signaling

to K). The same result was observed with trunk neural crest transfected with the early cranial-specific factors ( $n = 0/6$  embryos). Reprogrammed trunk neural crest, however, acquired chondrogenic potential and formed ectopic cartilage nodules ( $n = 4/7$  embryos) (Fig. 4, L to O) in the proximal jaw. Thus, introducing components of the cranial-specific transcriptional circuit is sufficient to reprogram trunk neural crest cells and to drive them to adopt an additional cartilaginous fate. These results definitively show that the cranial-specific regulatory circuit (Fig. 3J) we have defined confers chondrocytic potential to the trunk neural crest.

The development and differentiation of neural crest cells are controlled by a complex gene regulatory network, composed of transcription factors, signaling molecules, and epigenetic modifiers (12, 13). We have expanded the known cranial neural crest gene regulatory network by identifying transcriptional interactions specific to the cranial crest and absent from other subpopulations. By linking anterior identity in the gastrula to the expression of drivers of chondrocytic differentiation, we have identified a cranial-specific circuit (Fig. 3J) that endows the neural crest with its potential to differentiate into the craniofacial skeleton of vertebrates. Our results highlight how transcriptional circuits can be rewired to alter progenitor cell identity and fate during embryonic development.

#### REFERENCES AND NOTES

1. N. Le Douarin, *The Neural Crest* (Cambridge Univ. Press, 1982).
2. M. Simões-Costa, M. E. Bronner, *Genome Res.* **23**, 1069–1080 (2013).
3. T. Uesaka, M. Nagashima, H. Enomoto, *J. Neurosci.* **35**, 9879–9888 (2015).
4. V. Dyachuk *et al.*, *Science* **345**, 82–87 (2014).
5. C. S. Le Lièvre, N. M. Le Douarin, *J. Embryol. Exp. Morphol.* **34**, 125–154 (1975).
6. C. S. Le Lièvre, G. G. Schweitzer, C. M. Ziller, N. M. Le Douarin, *Dev. Biol.* **77**, 362–378 (1980).
7. P. Y. Lwigale, G. W. Conrad, M. Bronner-Fraser, *Development* **131**, 1979–1991 (2004).
8. P. Belancur, M. Bronner-Fraser, T. Sauka-Spengler, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3570–3575 (2010).
9. M. S. Simões-Costa, S. J. McKeown, J. Tan-Cabugao, T. Sauka-Spengler, M. E. Bronner, *PLoS Genet.* **8**, e1003142 (2012).
10. M. Simões-Costa, J. Tan-Cabugao, I. Antoshchukin, T. Sauka-Spengler, M. E. Bronner, *Genome Res.* **24**, 281–290 (2014).

#### ETHICS

## The social dilemma of autonomous vehicles

Jean-François Bonnefon,<sup>1</sup> Azim Shariff,<sup>2\*</sup> Iyad Rahwan<sup>3,†</sup>

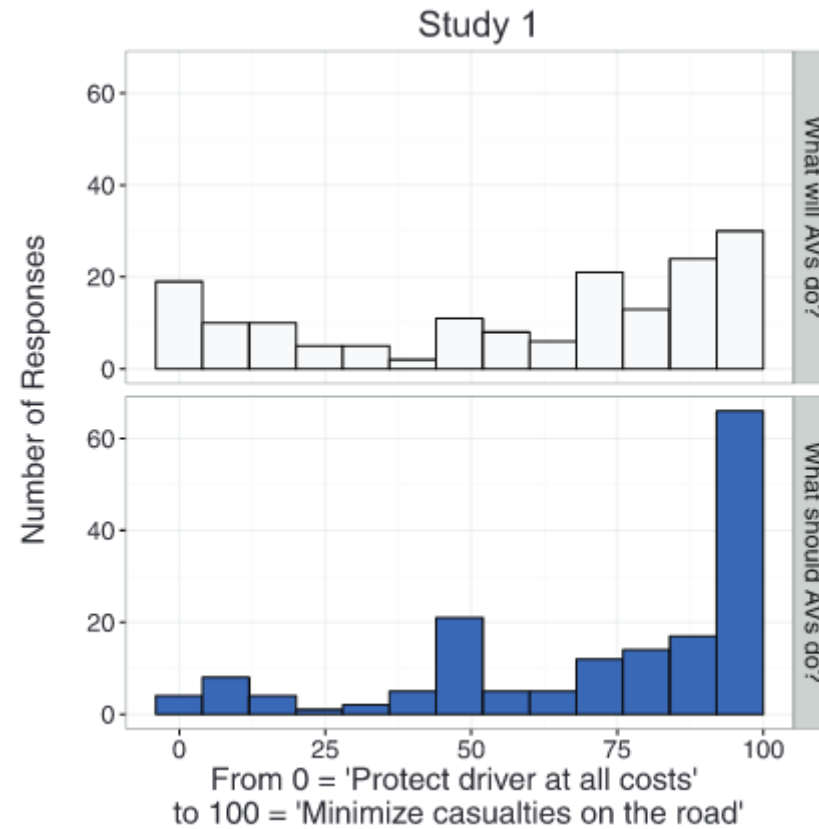
Autonomous vehicles (AVs) should reduce traffic accidents, but they will sometimes have to choose between two evils, such as running over pedestrians or sacrificing themselves and their passenger to save the pedestrians. Defining the algorithms that will help AVs make these moral decisions is a formidable challenge. We found that participants in six Amazon Mechanical Turk studies approved of utilitarian AVs (that is, AVs that sacrifice their passengers for the greater good) and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. The study participants disapprove of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV. Accordingly, regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology.

The year 2007 saw the completion of the first benchmark test for autonomous driving in realistic urban environments (1, 2). Since then, autonomous vehicles (AVs) such as Google's self-driving car covered thousands of miles of real-road driving (3). AVs have the potential to benefit the world by increasing traffic efficiency (4), reducing pollution (5), and eliminating up to 90% of traffic accidents (6). Not all crashes will be avoided, though, and some crashes will require AVs to make difficult ethical decisions in cases that involve unavoidable harm (7). For example, the AV may avoid harming several pedestrians by swerving and sacrificing a passerby, or the AV may be faced with the choice of sacrificing its own passenger to save one or more pedestrians (Fig. 1).

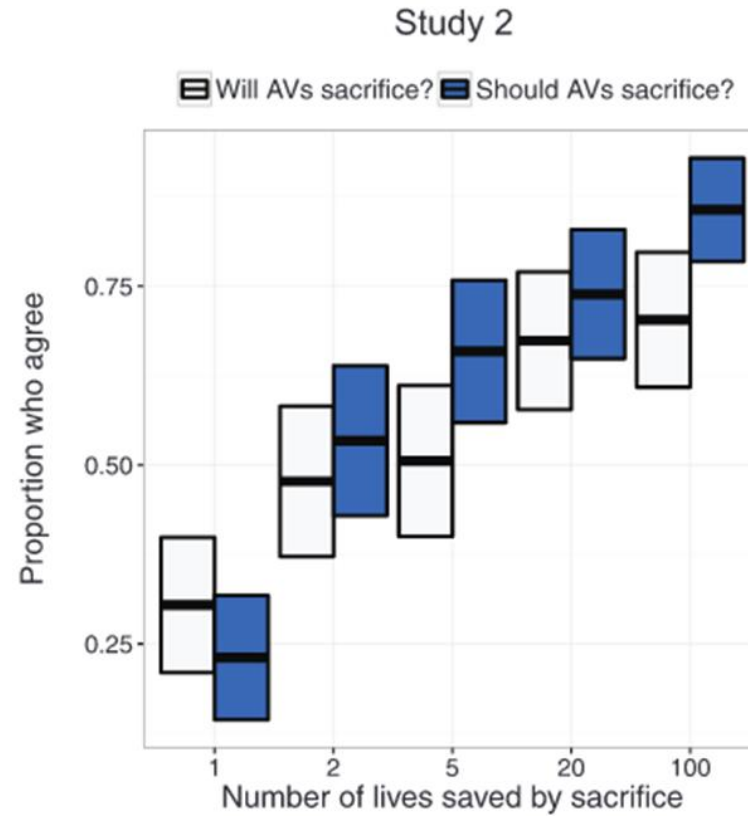
Although these scenarios appear unlikely, even low-probability events are bound to occur with millions of AVs on the road. Moreover, even if these situations were never to arise, AV programming must still include decision rules about what to do in such hypothetical situations. Thus, these types of decisions need be made well before AVs

the most common moral attitude is that the AV should swerve. This would fit a utilitarian moral doctrine (11), according to which the moral course of action is to minimize casualties. But consider then the case displayed in Fig. 1C. The utilitarian course of action, in that situation, would be for the AV to swerve and kill its passenger, but AVs programmed to follow this course of action might discourage buyers who believe their own safety should trump other considerations. Even though such situations may be exceedingly rare, their emotional saliency is likely to give them broad public exposure and a disproportionate weight in individual and public decisions about AVs. To align moral algorithms with human values, we must start a collective discussion about the ethics of AVs—that is, the moral algorithms that we are willing to accept as citizens and to be subjected to as car owners. Thus, we initiate the data-driven study of driverless car ethics, inspired by the methods of experimental ethics (12).

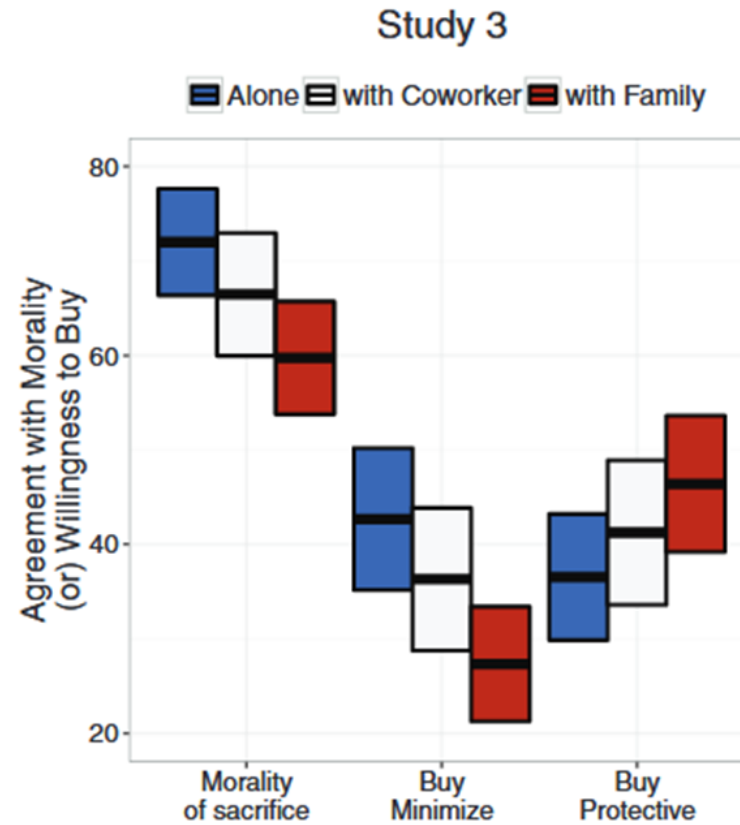
We conducted six online surveys ( $n = 1928$  total participants) between June and November 2015. All studies were programmed on Qualtrics survey



- Overwhelming moral preference for utilitarian AVs
- Future AVs will be less utilitarian than they should be



- Participants did not think AVs should sacrifice driver over only 1 pedestrian
- Moral approval of swerving increased with the number of lives saved



- Imagining that a family member was in the AV negatively affected the morality of sacrifice
- Despite overwhelming agreement with morality, the inclusion of other people shows increased preference for self-protective AVs



# Conclusions

- Regulating AVs may be necessary but also counterproductive
- Most people disapprove of a regulation that would enforce utilitarian AVs
- Such regulations could delay the adoption of AVs, negating the amount of lives saved by the utilitarian model

## Roadblock 2: Overreactions to inevitable accidents

### Problem

- Disproportionate news coverage may cause exaggerated danger perception
- Fear may worsen thorough *algorithm aversion*
- Deterrence of consumers and provocation of politicians into needless regulations

### Suggested Solutions

- Prepare the public for the inevitability of accidents
- Openly communicate algorithmic improvement
- Manage public overreaction with ‘fear placebos’ and information about actual risk levels

# A Tragic Loss

The Tesla Team • June 30, 2016

We learned yesterday evening that NHTSA is opening a preliminary evaluation into the performance of Autopilot during a recent fatal crash that occurred in a Model S. This is the first known fatality in just over 130 million miles where Autopilot was activated. Among all vehicles in the US, there is a fatality every 94 million miles. Worldwide, there is a fatality approximately every 60 million miles. It is important to emphasize that the NHTSA action is simply a preliminary evaluation to determine whether the system worked according to expectations.

Following our standard practice, Tesla informed NHTSA about the incident immediately after it occurred. What we know is that the vehicle was on a divided highway with Autopilot engaged when a tractor trailer drove across the highway perpendicular to the Model S. Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied. The high ride height of the trailer combined with its positioning across the road and the extremely rare circumstances of the impact caused the Model S to pass under the trailer, with the bottom of the trailer impacting the windshield of the Model S. Had the Model S impacted the front or rear of the trailer, even at high speed, its advanced crash safety system would likely have prevented serious injury as it has in numerous other similar incidents.

It is important to note that Tesla disables Autopilot by default and requires explicit acknowledgement that the system is new technology and still in a public beta phase before it can be enabled. When drivers activate Autopilot, the acknowledgment box explains, among other things, that Autopilot “is an assist feature that requires you to keep your hands on the steering wheel at all times,” and that “you need to maintain control and responsibility for your vehicle” while using it. Additionally, every time that Autopilot is engaged, the car reminds the driver to “Always keep your hands on the wheel. Be prepared to take over at any time.” The system also makes frequent checks to ensure that the driver’s hands remain on the wheel and provides visual and audible alerts if hands-on is not detected. It then gradually slows down the car until hands-on is detected again.

We do this to ensure that every time the feature is used, it is used as safely as possible. As more real-world miles accumulate and the software logic accounts for increasingly rare events, the probability of injury will keep decreasing. Autopilot is getting better all the time, but it is not perfect and still requires the driver to remain alert. Nonetheless, when

# A Tragic Loss

The Tesla Team · June 30, 2016

We learned yesterday evening that NHTSA is opening a preliminary evaluation into the performance of Autopilot during a recent fatal crash that occurred in a Model S. This is the first known fatality in just over 130 million miles where Autopilot was activated. Among all vehicles in the US, there is a fatality every 94 million miles. Worldwide, there is a fatality approximately every 60 million miles. It is important to emphasize that the NHTSA action is simply a preliminary evaluation to determine whether the system worked according to expectations.

Following our standard practice, Tesla informed NHTSA about the incident immediately after it occurred. What we know is that the vehicle was on a divided highway with Autopilot engaged when a tractor trailer drove across the highway perpendicular to the

Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied.

Images of the scene and circumstances of the impact caused the Model S to pass under the trailer, with the bottom of the trailer impacting the windshield of the Model S. Had the Model S impacted the front or rear of the trailer, even at high speed, its advanced crash safety system would likely have prevented serious injury as it has in numerous other similar incidents.

It is important to note that Tesla disables Autopilot by default and requires explicit acknowledgement that the system is new technology and still in a public beta phase before it can be enabled. When drivers activate Autopilot, the acknowledgement box explains, among other things, that Autopilot "is an assist feature that requires you to keep your hands on the steering wheel at all times," and that "you need to maintain control and responsibility for your vehicle" while using it. Additionally, every time that Autopilot is engaged, the car reminds the driver to "Always keep your hands on the wheel. Be prepared to take over at any time." The system also makes frequent checks to ensure that the driver's hands remain on the wheel and provides visual and audible alerts if hands-on is not detected. It then gradually slows down the car until hands-on is detected again.

We do this to ensure that every time the feature is used, it is used as safely as possible. As more real-world miles accumulate and the software logic accounts for increasingly rare events, the probability of injury will keep decreasing. Autopilot is getting better all the time, but it is not perfect, and still requires the driver to remain alert. Responsibility about

# Driver's family doesn't blame Tesla for fatal 'autopilot' crash

By Reuters September 11, 2017 | 3:34pm



Getty Images

The family of the driver of a Tesla Model S who was killed in a May 2016 crash while using the car's autonomous driving system said on Monday the was not to blame for the crash.

The statement from the family of Joshua Brown, released by a law firm, comes a day before the National Transportation Safety Board is set to hold hearing in Washington and vote on the probable cause of the crash.

"We heard numerous times that the car killed our son. That is simply not the case," said the statement.

MORE ON:  
**TESLA**

**Tesla prices secondary stock offering at discounted \$767 a share**

**Tesla discloses \$2 billion stock sale amid fresh SEC scrutiny**

**Musk reveals Cybertruck redesign was in the works if**

# Tesla driver killed while using autopilot was watching Harry Potter, witness says

Driver in first known fatal self-driving car crash was also driving so fast that 'he went so fast through my trailer I didn't see him', the truck driver involved said



▲ Investigation launched after Tesla driver killed while using autopilot

The Tesla driver killed in the first known fatal crash involving a self-driving car may have been watching a Harry Potter movie at the time of the collision in Florida, according to a truck driver involved in the crash.

The truck driver, Frank Baressi, 62, told the Associated Press that the Tesla driver Joshua Brown, 40, was "playing Harry Potter on the TV screen" during the collision and was driving so fast that "he went so fast through my trailer I didn't see him".

# Man killed in Tesla 'Autopilot' crash got numerous warnings: Report

PUBLISHED TUE, JUN 20 2017 7:25 AM EDT | UPDATED TUE, JUN 20 2017 10:10 AM EDT

REUTERS

SHARE f t in e




VIDEO 00:30  
**Man killed in Tesla 'Autopilot' crash reportedly got numerous warnings**

A man killed in a crash last year while using the semi-autonomous driving system on his Tesla Model S sedan kept his hands off the wheel for extended periods of time despite repeated automated warnings not to do so, a U.S. government report said on Monday.


The National Transportation Safety Board (NTSB) released 500 pages of findings into the May 2016 death of Joshua Brown, a former Navy SEAL, near Williston, Florida. Brown's Model S collided with a truck while it was engaged in the "Autopilot" mode and he was killed.

A Tesla spokeswoman Keely Sulprizio declined to comment on the NTSB report. In 2016, the company said Autopilot "does not allow the driver to abdicate responsibility," however.


## TRENDING NOW

 You should always keep a \$100 bill in your wallet, a psychologist says. Here's why experts agree

 Coronavirus updates: State Department will evacuate Americans from cruise ship, first death confirmed in Europe

 This 33-year-old paid off his \$300,000 house in 3 months—here's why he didn't invest the money

 Charlie Munger on Elon Musk: 'Never underestimate the man who overestimates himself'

 First coronavirus death confirmed in Europe, French health minister says



# Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation

Edmond Awad<sup>a,+</sup>, Sydney Levine<sup>a,b,c,+</sup>, Max Kleiman-Weiner<sup>b</sup>, Sohan Dsouza<sup>a</sup>, Joshua B. Tenenbaum<sup>b,\*</sup>, Azim Shariff<sup>d,\*</sup>, Jean-François Bonnefon<sup>e,\*</sup>, and Iyad Rahwan<sup>a,f,\*</sup>

<sup>a</sup>*Media Lab, Massachusetts Institute of Technology, MA, USA*

<sup>b</sup>*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, MA, USA*

<sup>c</sup>*Department of Psychology, Harvard University, MA, USA*

<sup>d</sup>*Department of Psychology and Social Behavior, University of California, Irvine, CA, USA*

<sup>e</sup>*Toulouse School of Economics (TSM-R), CNRS, University of Toulouse, France*

<sup>f</sup>*Institute for Data, Systems and Society, Massachusetts Institute of Technology, MA, USA*

<sup>+</sup>*Joint first author*

<sup>\*</sup>*Corresponding authors. e-mail: jbt@mit.edu; shariffa@uci.edu; jean-francois.bonnefon@tse-fr.eu; irahwan@mit.edu*

When a semi-autonomous car crashes and harms someone, how are blame and causal responsibility distributed across the human and machine drivers? In this article, we consider cases in which a pedestrian was hit and killed by a car being operated under shared control of a primary and a secondary driver. We find that when only one driver makes an error, that driver receives the blame and is considered causally responsible for the harm, regardless of whether that driver is a machine or a human. However, when both drivers make errors in cases of shared control between a human and a machine.











Driver	
1st	2nd
H	—
H	H
H	M
M	H
M	M
M	—

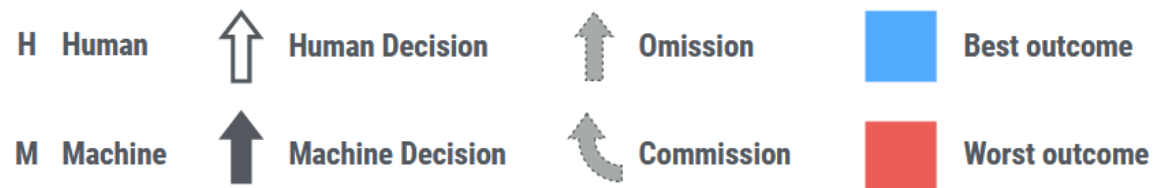
Driver		Real World Implementation
1st	2nd	
H	—	Regular
H	H	—
H	M	Guardian
M	H	Autopilot
M	M	—
M	—	Fully Automated



















*(Toyota Guardian)*

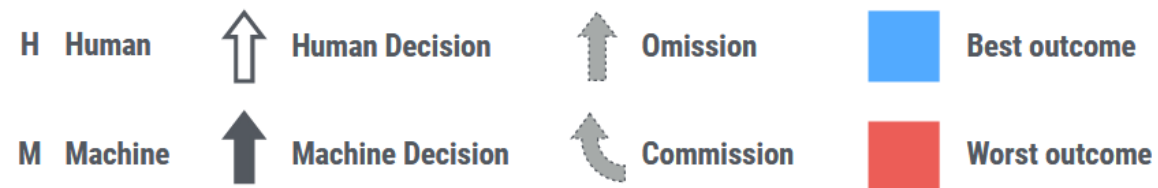
*(Tesla Autopilot)*

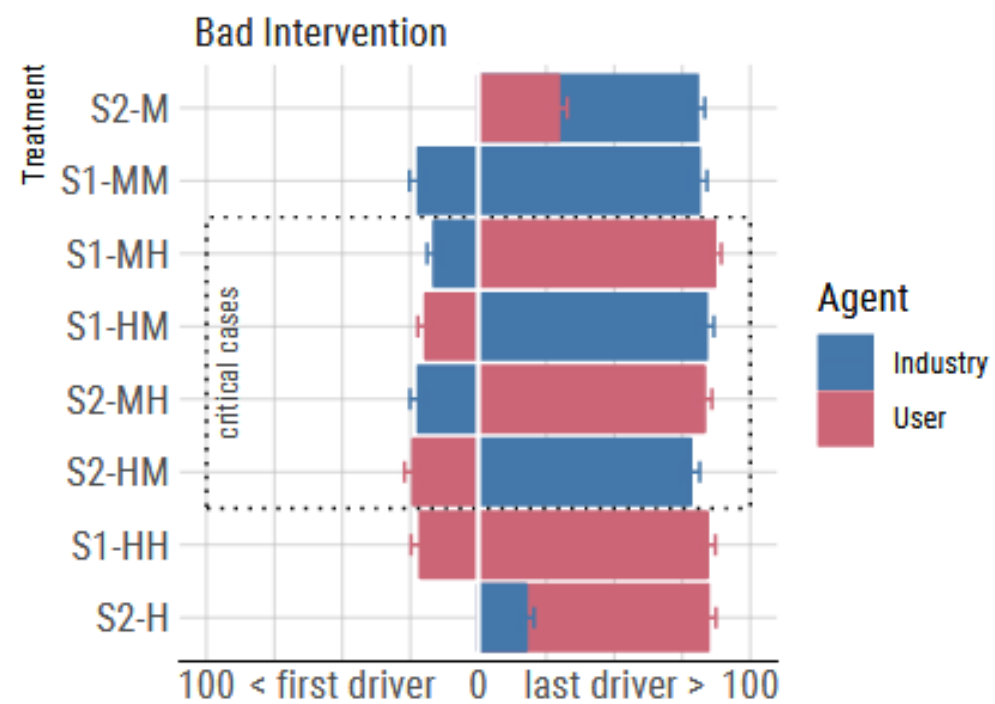


Driver		Real World Implementation	Bad Intervention
1st	2nd		
H	–	Regular	
H	H	–	 
H	M	Guardian	 
M	H	Autopilot	 
M	M	–	 
M	–	Fully Automated	



Driver		Real World Implementation	Bad Intervention	Missed Intervention
1st	2nd			
H	—	Regular		
H	H	—	 	 
H	M	Guardian	 	 
M	H	Autopilot	 	 
M	M	—	 	 
M	—	Fully Automated		







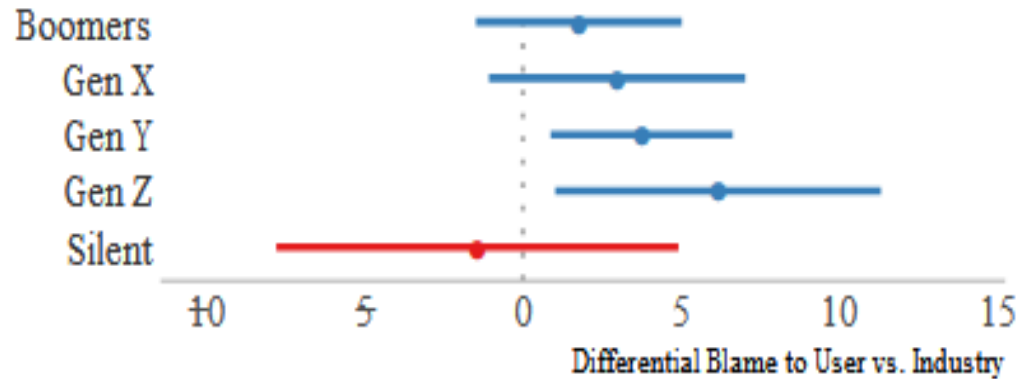
➡ 'man-in-the-loop' blamed more than 'machine-in-the-loop'

# Conclusions

- Less blame attributed to machine when both drivers make errors

Our central finding (diminished blame to the machine in dual-error cases) leads us to believe that, while there may be many psychological barriers to self-driving car adoption [19], public over-reaction to dual-error cases is not likely to be one of them. In fact, we should perhaps be concerned about public under-reaction.

- Trust in AVs increasing with each generation



## Roadblock 2: Overreactions to inevitable accidents

### Problem

- ~~• Disproportionate news coverage may cause exaggerated danger perception~~
- Bad intervention cases may spark disproportionate outrage
- Fear may worsen through *algorithm aversion*
- ~~• Deterrence of consumers and provocation of politicians into needless regulations~~
- Societal level responses may shape incentives for individual actors

### Suggested Solutions

- Prepare the public for the inevitability of accidents
- Openly communicate algorithmic improvement
- Manage public overreaction with ‘fear placebos’ and information about actual risk levels

## **Roadblock 3:** Lack of transparency in machine decision making process

### **Problem**

- Passengers will be accurately aware of car's rare failures but blissfully unaware of small successes and optimizations
- People can't comfortably predict and understand the behavior of the other entity
- Machine learning decision making is inherently opaque
- Inability to predict AV behavior will diminish trust in them

### **Suggested Solutions**

- Research the type of information required to form trustable mental models of autonomous vehicles.

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

Björn Meder,<sup>1,2,\*</sup> Nadine Fleischhut,<sup>3</sup> Nina-Carolin Krumnau,<sup>2</sup> and Michael R. Waldmann<sup>4</sup>

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

<sup>1</sup>MPRG iSearch, Max Planck Institute for Human Development, Berlin, Germany.

<sup>2</sup>Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany.

<sup>3</sup>Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany.

<sup>4</sup>Department of Psychology, University of Göttingen, Göttingen, Germany.



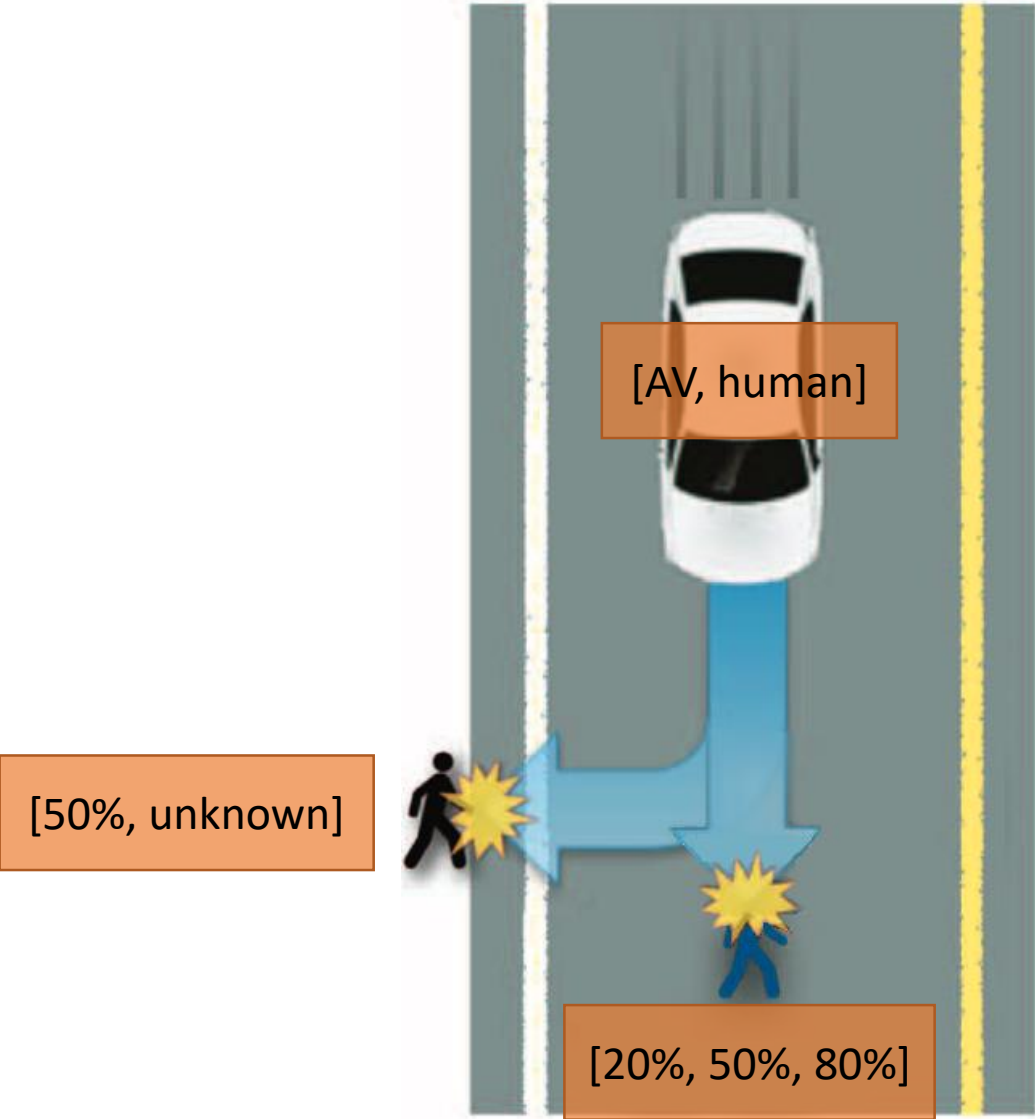
Is the trolley problem adequate?

```
if(moralDilemma) {  
    do_the_right_thing();  
}
```

# Is the trolley problem adequate?

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)

# Experiment 1



Dilemma

Someone is riding in a fully autonomous, self-driving car down a main road at the speed limit of 50 mph. In this self-driving car, an automated driving system performs all driving tasks under all conditions [Someone is driving down a main road at the speed limit of 50 mph]. Suddenly, a pedestrian appears on the street in front of the car. The car is equipped with a system that can estimate the emergency stopping distance based on the distance to the obstacle, the speed of the car, the road conditions, and the response time of the automated driving system [driver].

**The first option for the self-driving car [driver] is to stay in the lane and perform an emergency stop.** In this case, the car's systems estimate that the likelihood of colliding with the pedestrian is 50% [20% ; 80%]. This means that in 50 [20 ; 80] of 100 situations like this, a self-driving car [driver] cannot avoid a collision. In case of a collision, the pedestrian will certainly be seriously injured given the speed of the car.

**The second option for the self-driving car [driver] is to swerve to the right and perform an emergency stop.** Through the swerving maneuver, the car will avoid the collision with the pedestrian on the street. However, there is a bystander on the sidewalk to the right of the car. In this case, the car's systems are unable to estimate the likelihood of colliding with the bystander on the sidewalk: It might be lower, higher, or equal to that of colliding with the pedestrian on the street. [In this case, the car's systems estimate that the likelihood of colliding with the bystander on the sidewalk is 50%. This means that in 50 of 100 situations like this, a self-driving car [driver] cannot avoid a collision]. In case of collision, the bystander will certainly be seriously injured given the speed of the car.

The person in the car is not endangered by either maneuver, whether a collision occurs or not. There are no other vehicles behind the car, meaning that no road users besides the pedestrian on the street and the bystander on the sidewalk will be endangered by either maneuver. Swerving to the left is impossible due to oncoming traffic in the other lane.

Moral decision

Given the information provided, how should self-driving cars [drivers] behave in a situation like this?

☐ Stay in the lane and perform an emergency stop

☐ Swerve to the right and perform an emergency stop

Moral judgment

How morally acceptable is it for self-driving cars [drivers] to swerve to the right/stay in the lane and perform an emergency stop in a situation like this?

Completely unacceptable

☐ 1

☐ 2

☐ 3

☐ 4

☐ 5

☐ 6

Completely acceptable

Threshold value

We asked you how self-driving cars [drivers] should behave in a situation like this. When keeping the car in the lane and performing an emergency stop, there was a 50% [20% ; 80%] likelihood of colliding with the pedestrian on the street. When swerving to the right and performing an emergency stop, the likelihood of colliding with the bystander was uncertain [50%].

What should be the minimum likelihood of colliding with the pedestrian, in order to swerve to the right rather than staying in the lane?

☐ The likelihood of colliding with the pedestrian should at least be  % in order to swerve to the right rather than staying in the lane.

☐ Self-driving cars [drivers] should never swerve to the right, regardless of the likelihood of colliding with the pedestrian.

Probability estimate (uncertainty conditions only)

Please remember your previous decision, how self-driving cars [drivers] should behave in a situation like this. If the car swerved to the right and performed an emergency stop: What do/did you estimate to be the likelihood of colliding with the bystander on the sidewalk?

The likelihood of colliding with the bystander on the sidewalk is  %

When deciding how self-driving cars [drivers] should behave in a situation like this, did you consider how likely it was that the car would collide with the bystander on the sidewalk?

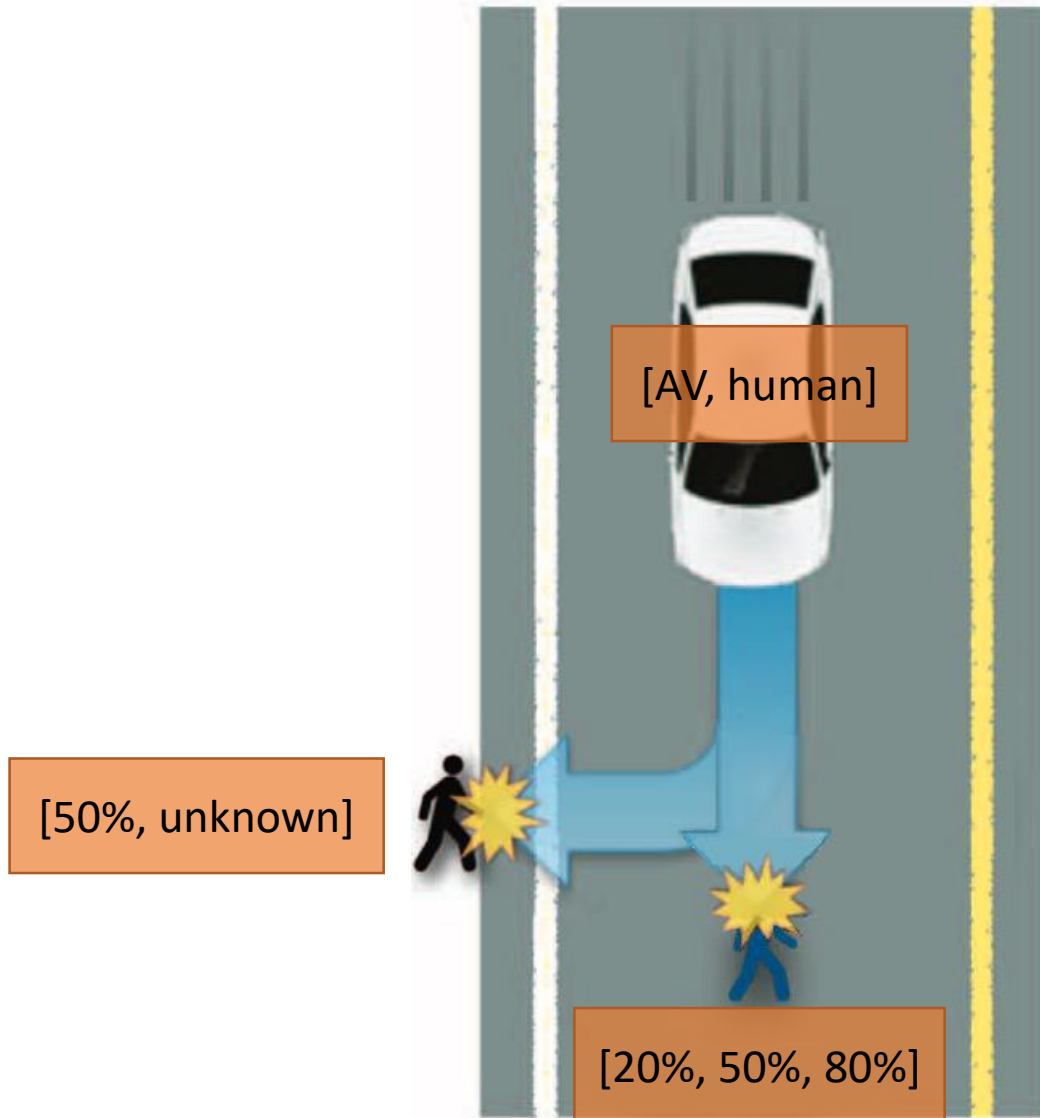
☐ Yes

☐ No

Decision rule

Manufacturers have to define how self-driving cars should behave in critical traffic situations [Driving teachers need to teach student drivers how they should behave in critical traffic situations]. Please try to formulate a rule on how self-driving cars [drivers] should behave in situations like the one presented in this study. Please also provide reasons for your answer.

# Experiment 1



**Q1: Decision Preference**

*(Should the car stay or swerve)*

**Q2: Moral Judgement**

*(How morally acceptable is it to stay or swerve?)*

**Q3: Swerving Threshold**

*(Minimum likelihood of collision to swerve)*

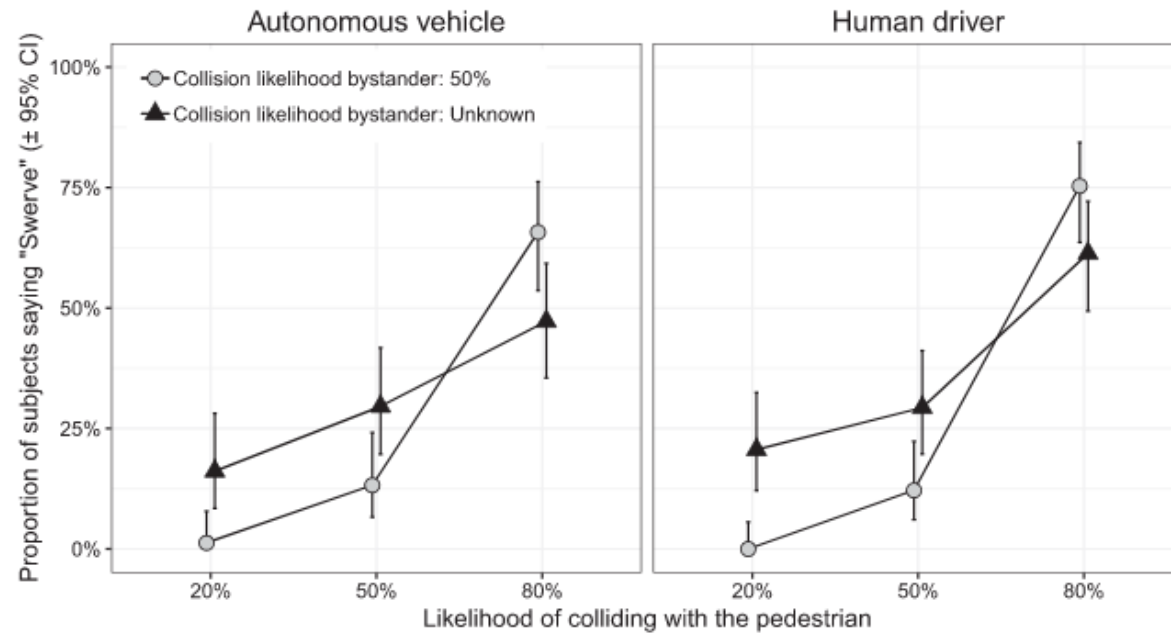
**Q4: Probability Estimate**

*(How likely is a collision with the bystander?)*

**Q5: Decision Rule**

*(What rule should the car follow in such a situation)*

# Results



## Q1: Decision Preference

*(Should the car stay or swerve)*

## Q2: Moral Judgement

*(How morally acceptable is it to stay or swerve?)*

## Q3: Swerving Threshold

*(Minimum likelihood of collision to swerve)*

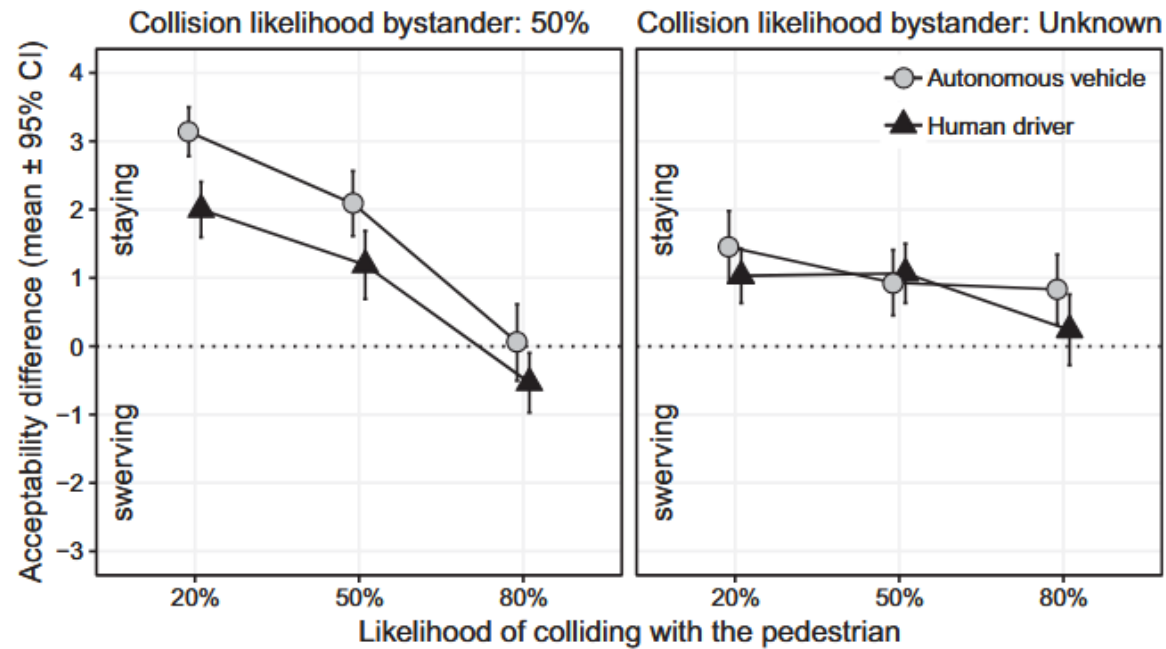
## Q4: Probability Estimate

*(How likely is a collision with the bystander?)*

## Q5: Decision Rule

*(What rule should the car follow in such a situation)*

# Results



Q1: Decision Preference

*(Should the car stay or swerve)*

Q2: Moral Judgement

*(How morally acceptable is it to stay or swerve?)*

Q3: Swerving Threshold

*(Minimum likelihood of collision to swerve)*

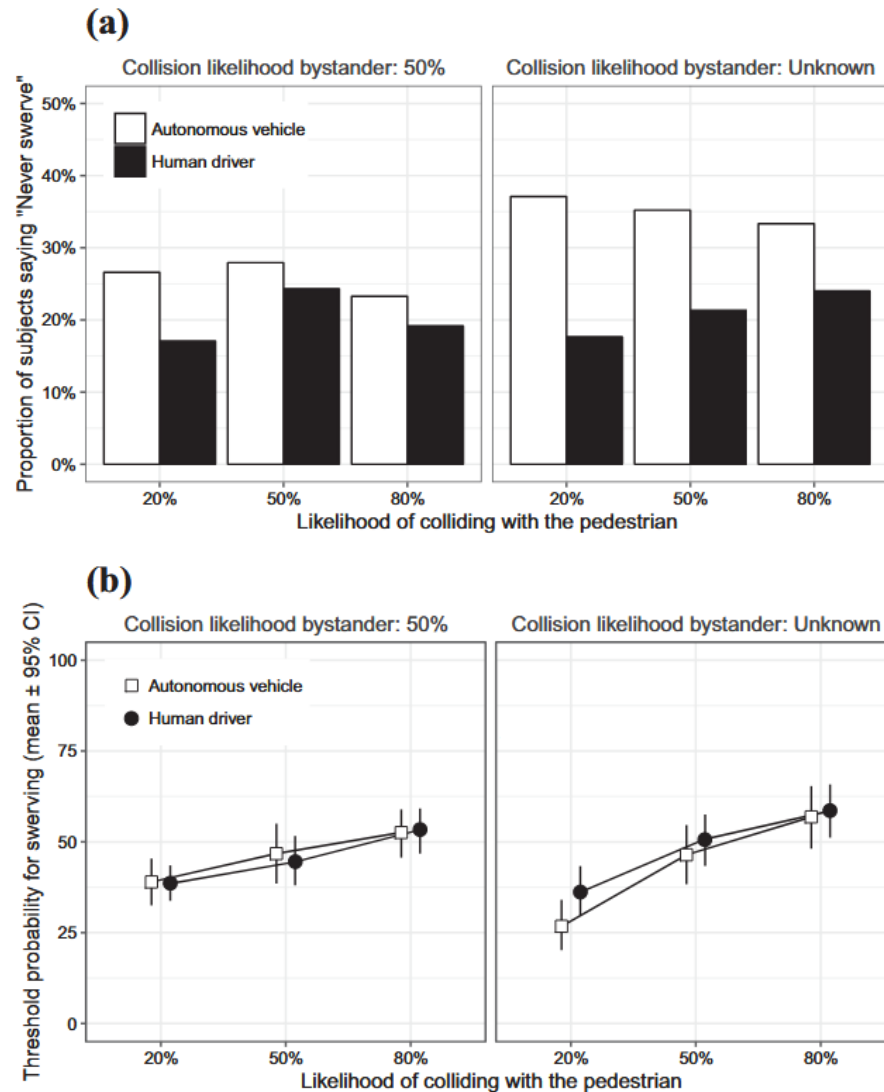
Q4: Probability Estimate

*(How likely is a collision with the bystander?)*

Q5: Decision Rule

*(What rule should the car follow in such a situation)*

# Results



Q1: Decision Preference

*(Should the car stay or swerve)*

Q2: Moral Judgement

*(How morally acceptable is it to stay or swerve?)*

Q3: Swerving Threshold

*(Minimum likelihood of collision to swerve)*

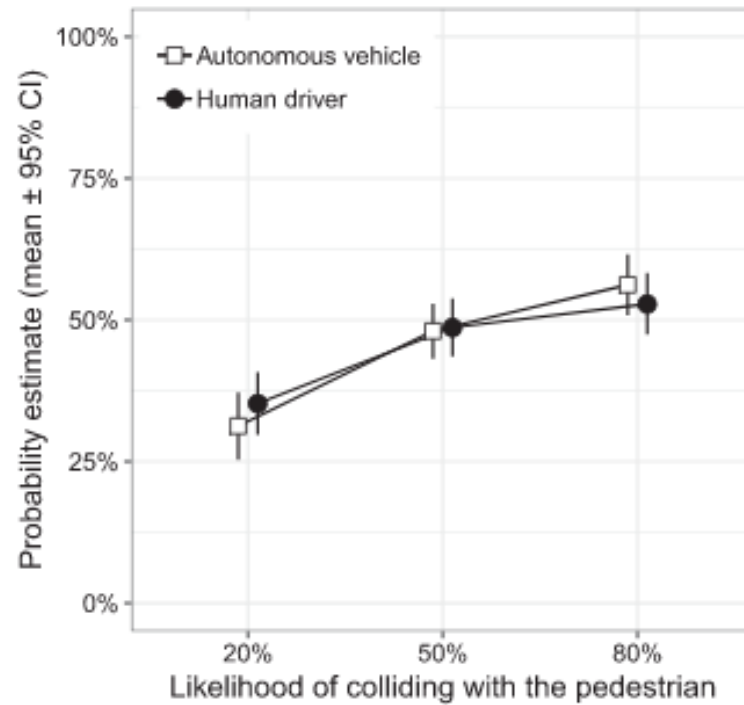
Q4: Probability Estimate

*(How likely is a collision with the bystander?)*

Q5: Decision Rule

*(What rule should the car follow in such a situation)*

# Results



## Q1: Decision Preference

*(Should the car stay or swerve)*

## Q2: Moral Judgement

*(How morally acceptable is it to stay or swerve?)*

## Q3: Swerving Threshold

*(Minimum likelihood of collision to swerve)*

## Q4: Probability Estimate

*(How likely is a collision with the bystander?)*

## Q5: Decision Rule

*(What rule should the car follow in such a situation)*



# Results

Description	Risk Bystander: 50%		Risk Bystander: Unknown	
	AV	HD	AV	HD
Always stay in the lane	18.7%	23.4%	19.5%	20.1%
Always swerve to the right	1.4%	2.7%	2.0%	3.8%
Only swerve if the likelihood of hitting the pedestrian is above a certain threshold/very high	2.4%	1.4%	5.0%	3.8%
Only swerve if the likelihood of hitting the bystander is below a certain threshold/very low	1.4%	2.7%	4.0%	1.0%
Only swerve if the likelihood of hitting the pedestrian is 100%	0.5%	1.8%	0.5%	1.0%
Only swerve if the likelihood of hitting the bystander is 0%	8.1%	4.5%	9.0%	9.1%
Always choose the option with the lowest (objective or subjective) likelihood of injuring somebody	34.4%	35.1%	14.0%	21.5%
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%
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%

## Q1: Decision Preference

*(Should the car stay or swerve)*

## Q2: Moral Judgement

*(How morally acceptable is it to stay or swerve?)*

## Q3: Swerving Threshold

*(Minimum likelihood of collision to swerve)*

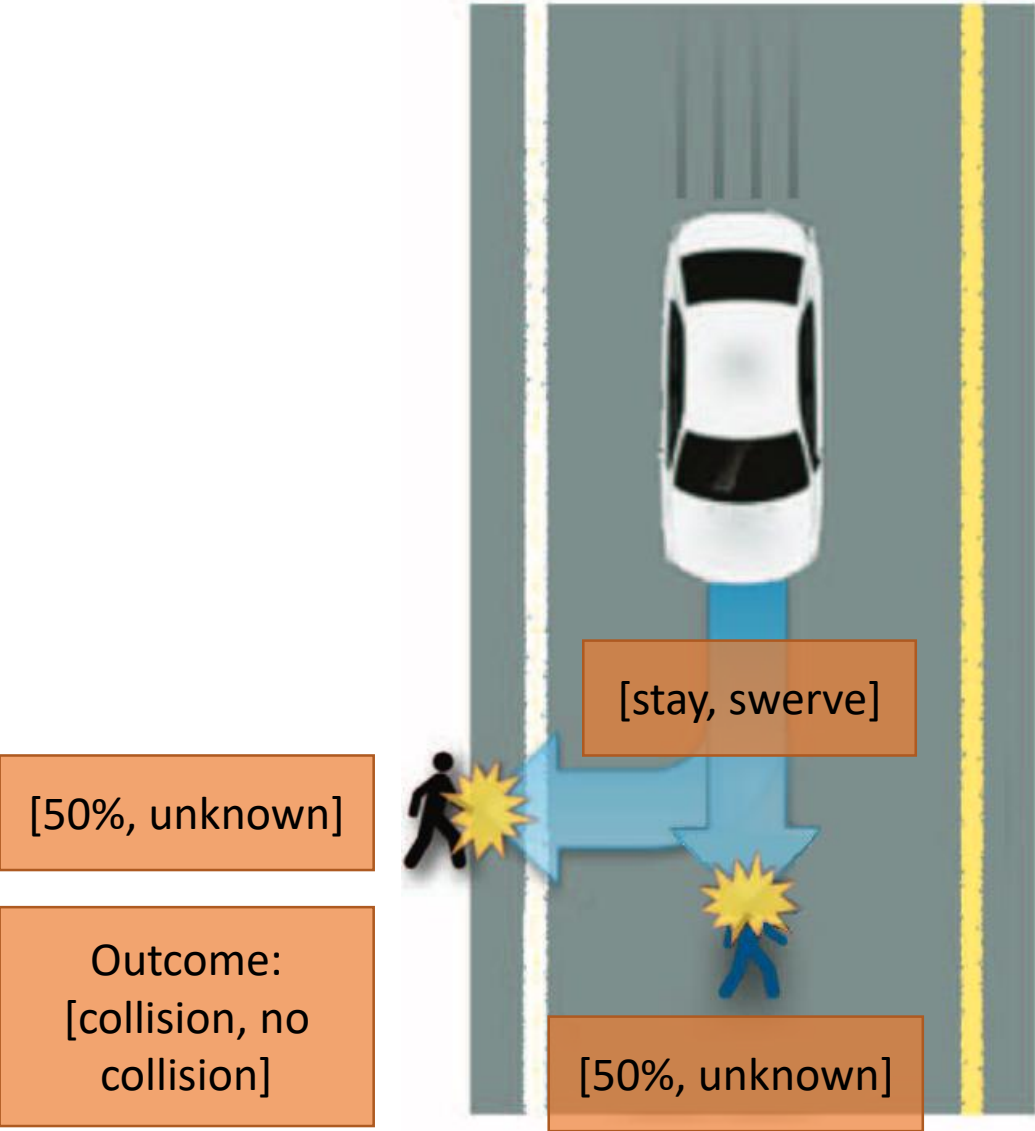
## Q4: Probability Estimate

*(How likely is a collision with the bystander?)*

## Q5: Decision Rule

*(What rule should the car follow in such a situation)*

# Experiment 2



Dilemma

Someone is riding in a fully autonomous, self-driving car down a main road at the speed limit of 50 mph. In this self-driving car, an automated driving system performs all driving tasks under all conditions. Suddenly, a pedestrian appears on the street in front of the car. The car is equipped with a system that can estimate the emergency stopping distance based on the distance to the obstacle, the speed of the car, the road conditions, and the response time of the automated driving system.

**The first option for the self-driving car is to stay in the lane and perform an emergency stop.**  
In this case, the car's systems estimate that the likelihood of colliding with the pedestrian on the street is 50% [the car's systems are unable to estimate the likelihood of colliding with the pedestrian on the street]. This means that in 50 of 100 situations like this, a self-driving car cannot avoid a collision. In case of a collision, the pedestrian will certainly be seriously injured given the speed of the car.

**The second option for the self-driving car is to swerve to the right and perform an emergency stop.**  
Through the swerving maneuver, the car will avoid the collision with the pedestrian on the street. However, there is a bystander on the sidewalk to the right of the car. In this case, the car's systems estimate that the likelihood of colliding with the bystander on the sidewalk is 50%. This means that in 50 of 100 situations like this, a self-driving car cannot avoid a collision. [the car's systems are unable to estimate the likelihood of colliding with the bystander on the sidewalk.] In case of a collision, the bystander will certainly be seriously injured given the speed of the car.

The person in the car is not endangered by either maneuver, whether a collision occurs or not. There are no other vehicles behind the car, meaning that no road users besides the pedestrian on the street and the bystander on the sidewalk will be endangered by either maneuver. Swerving to the left is impossible due to oncoming traffic in the other lane.

Good Outcome: No Collision

In the end, the self-driving car swerves to the right [stays in the lane] and performs an emergency stop. The car does not collide with the bystander on the sidewalk [pedestrian on the street] and no one is injured.

Bad Outcome: Collision

In the end, the self-driving car swerves to the right [stays in the lane] and performs an emergency stop. The car collides with the bystander on the sidewalk [pedestrian on the street] and the bystander [pedestrian] is seriously injured.

Moral decision

How should self-driving cars behave in a situation like this? Please answer as if you would not know the actual decision and outcome!

☐ Stay in the lane and perform an emergency stop

☐ Swerve to the right and perform an emergency stop

Moral judgment

How morally acceptable is it for self-driving cars to swerve to the right/stay in the lane and perform an emergency stop in a situation like this? Please answer as if you would not know the actual decision and outcome!

Completely unacceptable

☐ 1

☐ 2

☐ 3

☐ 4

☐ 5

☐ 6

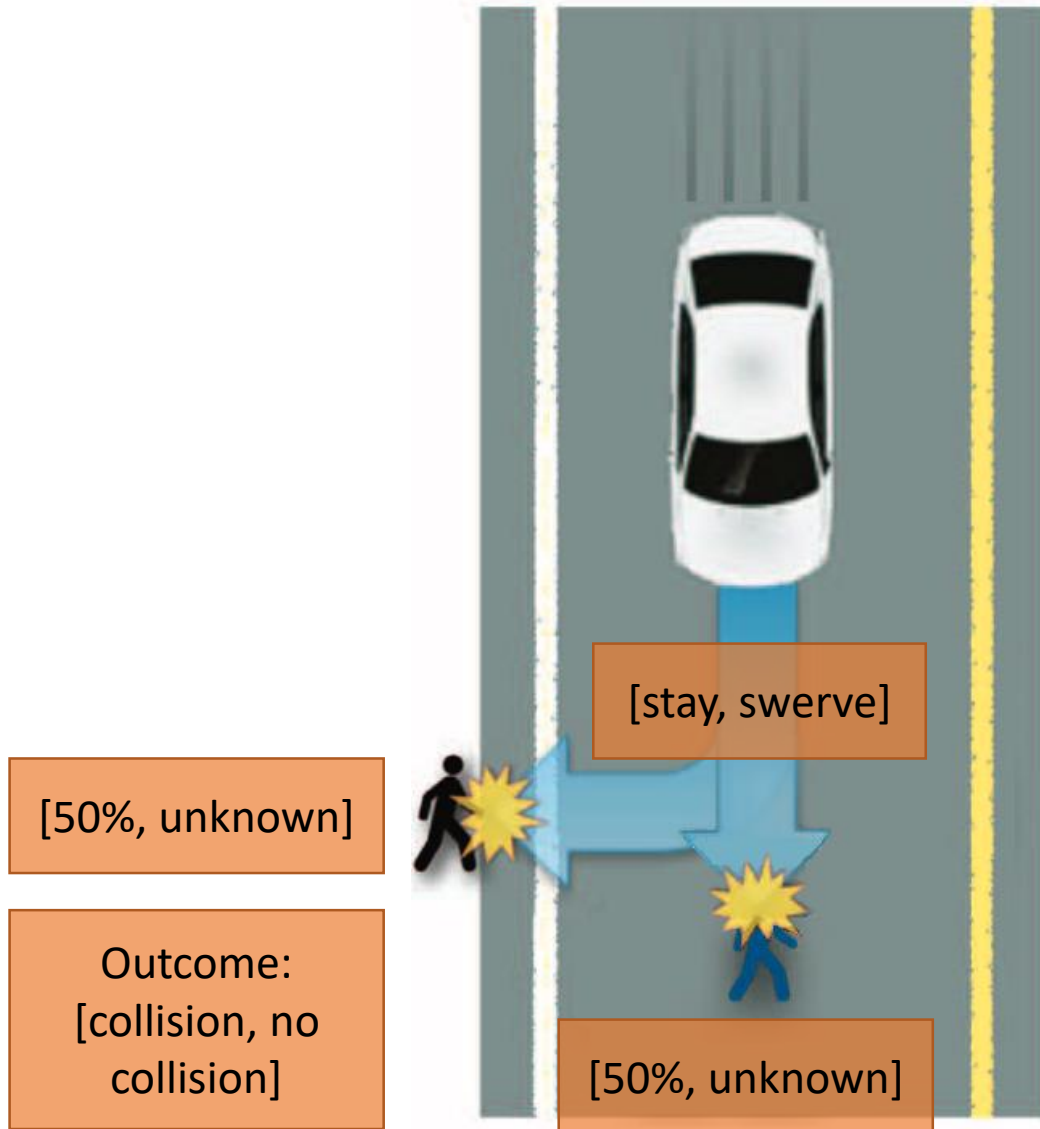
Completely acceptable

Probability estimate (uncertainty conditions only)

If the car swerved to the right and performed an emergency stop: What do/did you estimate to be the likelihood of colliding with the bystander on the sidewalk?  
Please answer as if you would not know the actual decision and outcome!  
The likelihood of colliding with the bystander on the sidewalk is/was  %

If the car stayed in the lane and performed an emergency stop: What do/did you estimate to be the likelihood of colliding with the pedestrian on the street?  
Please answer as if you would not know the actual decision and outcome!  
The likelihood of colliding with the pedestrian on the street is/was  %

## Experiment 2



**Q1: Decision Preference**

*(Should the car stay or swerve)*

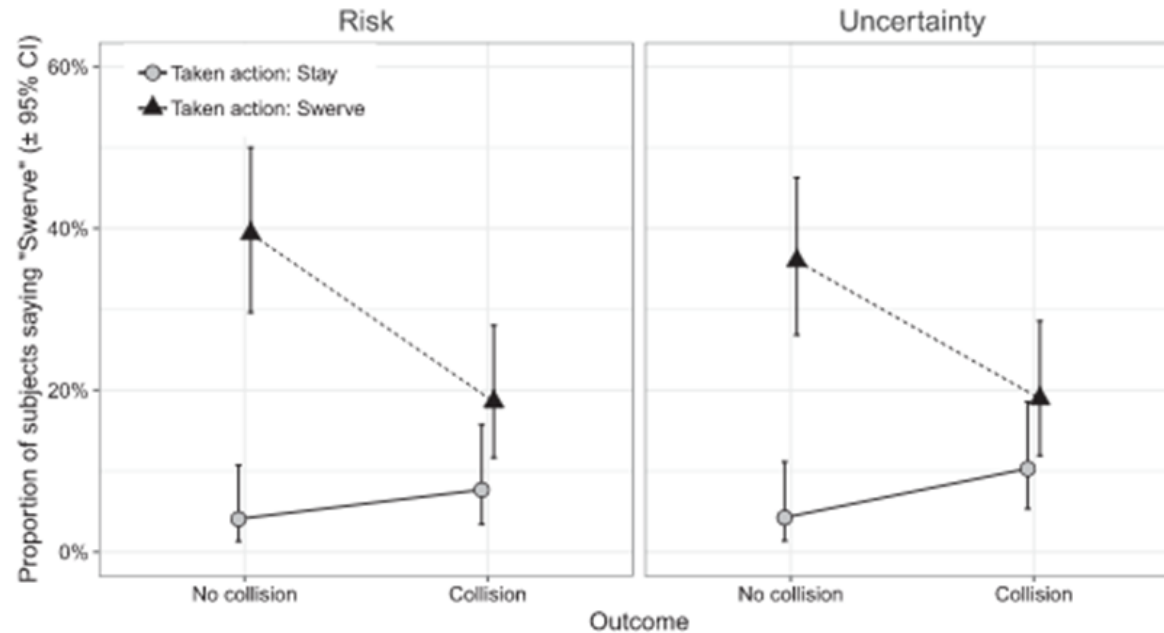
**Q2: Moral Judgement**

*(How morally acceptable is it to stay or swerve?)*

**Q4: Probability Estimate**

*(How likely is a collision with the bystander?)*

# Results



## Q1: Decision Preference

*(Should the car stay or swerve)*

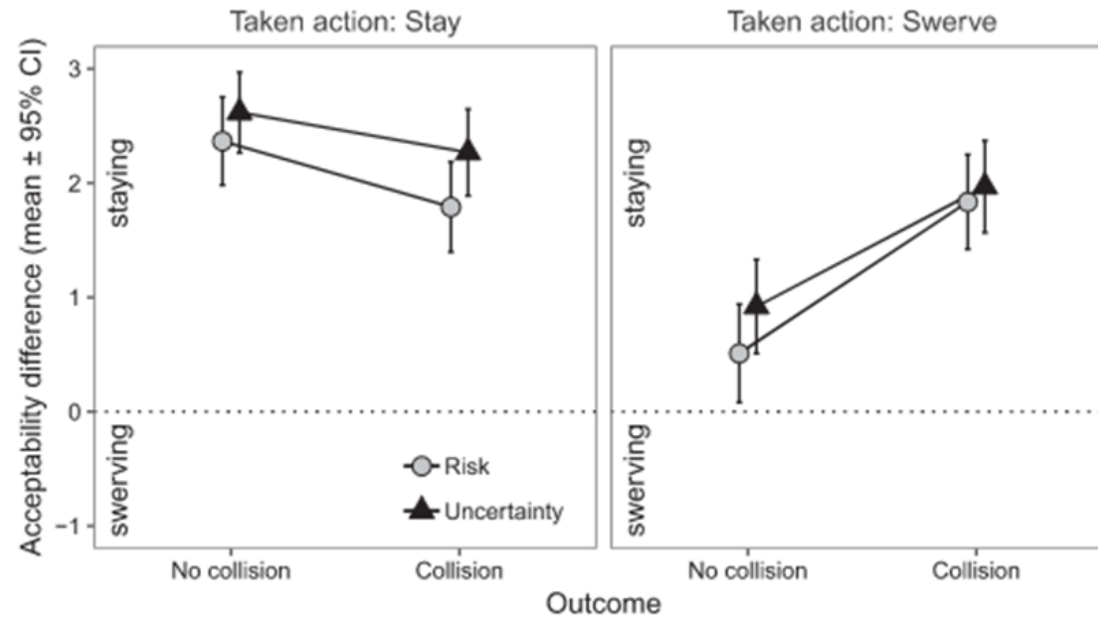
## Q2: Moral Judgement

*(How morally acceptable is it to stay or swerve?)*

## Q4: Probability Estimate

*(How likely is a collision with the bystander?)*

# Results



Q1: Decision Preference

*(Should the car stay or swerve)*

Q2: Moral Judgement

*(How morally acceptable is it to stay or swerve?)*

Q4: Probability Estimate

*(How likely is a collision with the bystander?)*

# Conclusions

- Dilemma scenarios should reflect scenarios AVs actually undergo
- People respond differently given the degree of uncertainty
- General preference for staying in lane

## **Utilitarian consequentialist:**

*The moral act is the one that statistically maximizes some social utility criterion*

## **Rule consequentialist:**

*The rightness of an action derives from whether the action maximizes utility in a class of situations governed by a rule*

## Roadblock 3: Lack of transparency in machine decision making process

### Problem

- Passengers will be accurately aware of car's rare failures but blissfully unaware of small successes and optimizations
- People can't comfortably predict and understand the behavior of the other entity
- Machine learning decision making is inherently opaque
- Inability to predict AV behavior will diminish trust in them

### Suggested Solutions

- ~~Research the type of information required to form trustable mental models of autonomous vehicles.~~
- A rule-consequentialist driving model may make AV driving patterns familiar and predictable