nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0762-8

Drivers are blamed more than their automated cars when both make mistakes

Edmond Awad[1,2,9], Sydney Levine[1,3,4,9], Max Kleiman-Weiner[3,4], Sohan Dsouza[1], Joshua B. Tenenbaum[3]*, Azim Shariff[5]*, Jean-François Bonnefon[1,6]* and Iyad Rahwan[1,7,8]*

When an automated car harms someone, who is blamed by those who hear about it? Here we asked human participants to consider hypothetical cases in which a pedestrian was killed by a car operated under shared control of a primary and a secondary driver and to indicate how blame should be allocated. We find that when only one driver makes an error, that driver is blamed more regardless of whether that driver is a machine or a human. However, when both drivers make errors in cases of human–machine shared-control vehicles, the blame attributed to the machine is reduced. This finding portends a public under-reaction to the malfunctioning artificial intelligence components of automated cars and therefore has a direct policy implication: allowing the de facto standards for shared-control vehicles to be established in courts by the jury system could fail to properly regulate the safety of those vehicles; instead, a top-down scheme (through federal laws) may be called for.

Every year, about 1.25 million people die worldwide in car crashes[1]. Laws concerning principles of negligence currently adjudicate how responsibility and blame are assigned to the individuals who injure others in these harmful crashes. The impending transition to fully automated cars promises a radical shift in how blame and responsibility will be attributed in the cases where crashes do occur, but most agree that little or no blame will be attributed to the occupants in the car who will, by then, be entirely removed from the decision-making loop[2]. However, before this era of fully automated cars arrives, we are entering a delicate era of shared control between humans and machines.

This new moment signals a departure from our current system—where individuals have full control over their vehicles and thereby bear full responsibility for crashes (absent mitigating circumstances)—to a new system where blame and responsibility may be shared between a human and a machine driver. The spontaneous reactions of people to crashes that occur when a human and machine share control of a vehicle have at least two direct industry-shaping implications. First, at present, little is known about how the public is likely to respond to crashes that involve both human and machine drivers. This uncertainty has concrete implications: manufacturers price products to reflect the liability they expect to incur from the sale of those products. If manufacturers cannot assess the scope of the liability they will incur from automated vehicles, that uncertainty will translate to substantially inflated prices of automated vehicles[2]. Moreover, the rate of the adoption of automated vehicles will be proportional to the cost to consumers in adopting the new technology[2]. (The rate of the adoption of this technology is contingent on many other factors, including consumers' understanding of the relative risks and benefits of using the cars. We do not mean to state that uncertainty about the scope of liability for manufacturers is the only factor impacting adoption, just that it is an important one.) Accordingly, the uncertainty about the extent of corporate liability for automated vehicle crashes may be slowing down automated vehicle adoption[2] while people continue to die in car crashes each year. Clarifying how and when responsibility will be attributed to manufacturers in automated car crashes will be a first step in reducing this uncertainty and speeding the adoption of automated, and eventually fully automated, vehicles.

The second direct implication of this work will be to forecast how a tort-based regulatory scheme (which is decided on the basis on jury decisions) is likely to turn out. Put another way, understanding how the public is likely to react to crashes that involve both a human and a machine driver will give us a hint as to what standards will be established if we let jury decisions shape them. If our work uncovers systematic biases that are likely to impact juries and would impede the adoption of automated cars, then it may make sense for federal regulations be put in place, which would pre-empt the tort system from being the avenue for establishing standards for these cars.

Already, automated vehicle crashes are in the public eye. In May 2016, the first deadly crash of a Tesla Autopilot car occurred and the occupant of the car was killed. In a news release, Tesla explained: "Neither Autopilot nor the driver noticed the white side of the tractor-trailer against a brightly lit sky, so the brake was not applied"[3]. In March 2018, the first automated car crash that killed a pedestrian occurred. A pedestrian that was crossing the street went unnoticed by both the car and the back-up driver (Uber). A few seconds before the crash, the car finally identified that it should be braking but failed to do so. The driver also braked too late to avoid the collision.

In the fatal Tesla and Uber crashes, both the machine driver and the human driver should have taken action and neither did. The mistakes of both the machine and the human led to the crash. The National Highway Safety Traffic Administration carried out an investigation of the Tesla incident and did not find Tesla at fault

[1]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Economics, University of Exeter Business School, Exeter, UK. [3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Department of Psychology, Harvard University, Cambridge, MA, USA. [5]Department of Psychology, University of British Columbia Vancouver, Vancouver, British Columbia, Canada. [6]Toulouse School of Economics (TSM-Research), Centre National de la Recherche Scientifique, University of Toulouse Capitole, Toulouse, France. [7]Centre for Humans & Machines, Max-Planck Institute for Human Development, Berlin, Germany. [8]Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, MA, USA. [9]These authors contributed equally: Edmond Awad and Sydney Levine. *e-mail: jbt@mit.edu; shariffa@uci.edu; jean-francois.bonnefon@tse-fr.eu; irahwan@mit.edu

in the crash[4]. Likewise, Uber has been exonerated from criminal charges after an investigation by a county prosecutor[5]. Notably, press attention surrounding the Tesla incident was markedly skewed towards blaming the human driver for the crash, with rumours quickly circulating that the driver had been watching a Harry Potter movie[6], though following further investigation it was discovered that there was no evidence grounding this claim[7]. Likewise, the fate of the Uber back-up driver remains unknown[5], with press attention focusing on the distracted nature of the driver[8].

The set of anecdotes around these two crashes begins to suggest a troubling pattern, namely that humans might be blamed more than their machine partners in certain kinds of automated vehicle crashes. Was this pattern a fluke of the circumstances of the crash and the press environment? Or does it reflect something psychologically deeper that may colour our responses to human–machine joint action and, in particular, when a human–machine pair jointly controls a vehicle?

What we are currently witnessing is a gradual and multi-pronged increase toward full automation, going through several steps of shared control between user and vehicle, which may take decades due to technical and regulatory issues as well as the attitudes of consumers towards adoption[9,10] (see Fig. 1). Some vehicles can take control over the actions of a human driver (for example, Toyota's Guardian) to perform emergency manoeuvres. Other vehicles may do most of the driving, while requiring the user to constantly monitor the situation and be ready to take control (for example, Tesla's Autopilot). Unless clear or explicitly mentioned, we use 'human' and 'user' interchangeably to refer to the person inside the car (being a driver or a passenger), and we use 'industry' and 'machine' interchangeably to refer to both company and car combined.

Our central question is this: when an automated car crashes and harms someone, how is blame and causal responsibility attributed to the human and machine drivers by people who hear about the crash? In this article, we use vignettes in which a pedestrian was hit and killed by a car being operated under shared control of a primary and a secondary driver, and ask our participants to evaluate the crash on metrics of blame and causal responsibility. The cases we use are hypothetical (insofar as respondents know that they did not actually take place), but are not unrealistic as they were designed to contain the relevant elements of events that could actually occur. We consider a wide range of control regimes (see Fig. 1), but the two main cases of interest are the instances of shared control where a human is the primary driver and the machine a secondary driver (human–machine) and where the machine is the primary driver and the human the secondary driver (machine–human). We consider a simplified space of scenarios in which (1) the main driver makes the correct choice and the secondary driver incorrectly intervenes (bad intervention) and (2) the main driver makes an error and the secondary driver fails to intervene (missed intervention). Both scenarios end in a crash. For comparison, we also include analogous scenarios involving a single human driver (a regular car) or a single machine driver (a fully automated car) as well as two hypothetical two-driver cars (driven by two humans or two machines). We ask participants to make evaluations of the human user and one representative of the machine, either the car itself or the company that designed the car.

In bad intervention cases (see Case description for details), the primary driver (be it human or machine) has made a correct decision to keep the car on course, which will avoid a pedestrian. Following this, the secondary driver makes the decision to swerve the car into the pedestrian. In these sorts of cases, we expect that the secondary driver (the only driver that makes a mistake) will be blamed more than the first driver. What is less clear is whether people will assign blame and causal responsibility differently if this secondary driver is a human driver or a machine. Recent research suggests that humans may be blamed more than robots for making the same error in the same situations[11].

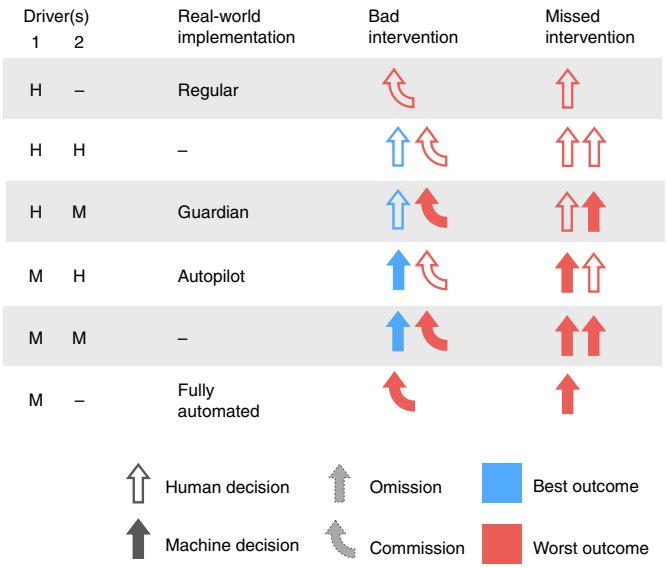

Fig. 1 | Actions or action sequences for the different car types considered. Outline arrows indicate an action by a human (H) and solid arrows indicate an action by a machine (M). The top and bottom rows represent sole-driver cars, while all others represent dual-driver cars. A red arrow indicates a decision—whether action or inaction—that had the avoidable death of a pedestrian as the outcome; a blue arrow indicates a decision that does not result in any deaths. For example, the H + M type (real-world implementation is the Guardian system) has a human main driver (1) and a machine standby driver (2). A bad intervention then involves the human staying on course (a non-lethal action, indicated by the outline of the straight, blue arrow) and the machine overriding that action, causing the death of the pedestrian (solid, angled red arrow). A missed intervention involves the human staying on course to kill the pedestrian (outline, straight red arrow) without intervention from the machine (solid, straight red arrow).

In missed intervention cases, the primary driver has made an incorrect decision to keep the car on course (rather than swerving), which would cause the car to hit and kill a pedestrian. The secondary driver then neglects to swerve out of the way of the pedestrian. In these cases, the predictions for how participants will distribute blame and causal responsibility are less clear because both drivers make a mistake. As in the bad intervention cases, agent type (human or machine) may have an effect on blame and causal responsibility ratings. But, unlike with bad intervention cases, missed intervention cases introduce the possibility that driver role (primary or secondary) may also impact judgements. It is possible that participants may shift responsibility and blame either toward the agent who contributed the most to the outcome (primary driver) or to the agent who had the last opportunity to act (secondary driver[12–15]). Under some regimes—such as Toyota's Guardian—the user does most of the driving but the decision to override (and thus to act last) pertains to the machine. Under others—such as Tesla's Autopilot—the machine does most of the driving but the decision to override pertains to the user.

Results

All studies used hypothetical vignettes that describe a crash (see Case description for details on car regimes and intervention types, and see Supplementary methods 1 for vignettes of Studies 1–5).

Study 1. Study 1 compared four kinds of car with different regimes of control. Each car had a primary driver whose job it was to drive the car, and a secondary driver whose job it was to monitor the actions of

Table 1 | Regression analysis of data collected in Studies 1–5 in the cases of bad intervention and missed intervention

	Blame and causal responsibility							
	Bad intervention			Missed intervention				
	Study 1	Study 2	Study 3	Study 1	Study 2	Study 3	Study 4	Study 5
Human	2.141	3.358	−1.508	**16.942**	**17.493**	**3.567**	**10.745**	**2.594**
	(1.061)	(1.574)	(0.811)	**(1.148)**	**(1.514)**	**(0.852)**	**(2.189)**	**(0.860)**
	$P=0.044$	$P=0.033$	$P=0.063$	$P=0.000$	$P=0.000$	$P=0.000$	$P=0.000$	$P=0.003$
Mistake	**64.293**	**57.559**	**11.917**					
	(1.061)	**(1.574)**	**(0.881)**					
	$P=0.000$	$P=0.000$	$P=0.000$					
Last driver				−1.822	−6.759	1.715	−0.073	1.355
				(0.915)	(1.514)	(0.852)	(2.189)	(0.860)
				$P=0.047$	$P=0.000$	$P=0.045$	$P=0.974$	$P=0.116$
Constant	18.653	21.352	27.406	60.504	57.354	36.102	61.032	65.923
	(0.916)	(1.370)	(1.171)	(1.531)	(1.878)	(1.252)	(1.911)	(0.811)
	$P=0.000$	$P=0.000$	$P=0.000$	$P=0.000$	$P=0.000$	$P=0.000$	$P=0.000$	$P=0.000$
Participant random effects?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Question random effects?	Yes	Yes	Yes	Yes	Yes	Yes	N/A	N/A
n	786	382	389	786	382	389	375	2,000
Observations	3,144	1,528	3,112	3,144	1,528	3,112	750	4,000

Data from Studies 2 and 3 are limited to shared-control regimes in the table. 'Human' refers to the type of agent in question (that is, human as compared to the baseline, machine), 'Mistake' refers to whether the decision was a mistake (that is, the decision would have resulted in losing a life or losing more lives in Study 3) and 'Last driver' refers to the driver role (that is, the driver assumes the secondary role). All models include participant random effects and question (blame or causal responsibility) random effects, where applicable. Data were assumed to meet the requirements of the model. N/A, not applicable.

the first driver and intervene when the first driver made an error. The car architectures of central interest were human primary–machine secondary (human–machine) and machine primary–human secondary (machine–human). We also included human–human and machine–machine architectures for comparison. This allowed us to see how blame was distributed in a dual-driver architecture when there was no difference in driver type (human or machine) in each of the driving roles (primary or secondary).

Bad interventions. In bad intervention cases, two predictors were entered into a regression with blame and causal responsibility ratings as the outcome variable: (1) whether or not the driver made an error and (2) driver type (human or machine). The main finding is that whether or not the driver made an error was a significant predictor of ratings (see Table 1, bad intervention, Study 1). In other words, unnecessary intervention by a driver leading to the death of a pedestrian was blamed more than a driver that operated on the correct course—regardless of whether the driver was a human or machine. It is worth noting here that we did not detect a reliable effect of driver type (human versus machine), once correcting for multiple comparisons (see Table 1, human, bad intervention, Study 1). We do not discuss this factor further in the bad intervention cases.

Missed interventions. In missed intervention cases, blame and responsibility judgements cannot depend on whether a driver made an error because both drivers make errors in these cases. The main finding for these cases is that driver type—whether the driver is a human or machine—has a significant impact on ratings. Specifically, in these shared-control scenarios where both human and machine have made errors, the machine driver is consistently blamed less than the human driver (Table 1, missed intervention, Study 1) and Fig. 2).

The human–machine difference appears to be driven by a reduction in the blame attributed to machines when there is a human in the loop. This is evident when comparing both the human–machine and machine–human instances of shared control to the machine–machine scenario. Note that the behaviours in these scenarios are identical, but the extent to which a machine is blamed depends on whether it is sharing control with a human or operating both the primary and secondary driver roles. When the machine is the primary driver, it is held significantly less blameworthy when its secondary driver is a human ($m_1=57.2$) compared to when the secondary driver is also the machine ($m_2=68$), $t(760.6)=-5.0$, $P<0.0001$, $m_2-m_1=10.8$, 95% CI for $m_2-m_1=6.6$–15 (all tests are two-tailed.) Similarly, when the machine is the secondary driver, it is held significantly less blameworthy when its primary driver is a human ($m_1=53.4$) compared to when the primary driver is also the machine ($m_2=68$), $t(722.77)=-6.6$, $P<0.0001$, $m_2-m_1=14.6$, 95% CI for $m_2-m_1=10.2$–19.

Study 2. Study 2 compared the human–machine and machine–human shared control cars to two different baseline cars: a standard car exclusively driven by a human and a fully automated car exclusively driven by a machine. This allowed us to both replicate the main results of Study 1 (the responses to machine–human and human–machine crashes) and determine how blame was assigned differently to dual-driver cars as compared to sole-driver cars. The industry representative was varied (car and company), but this exploratory variable was analysed neither in this study nor in subsequent studies.

Bad interventions. We replicated the main results of Study 1: namely, in bad intervention cases for the shared-control cars (machine–human and human–machine), whether or not the driver made an

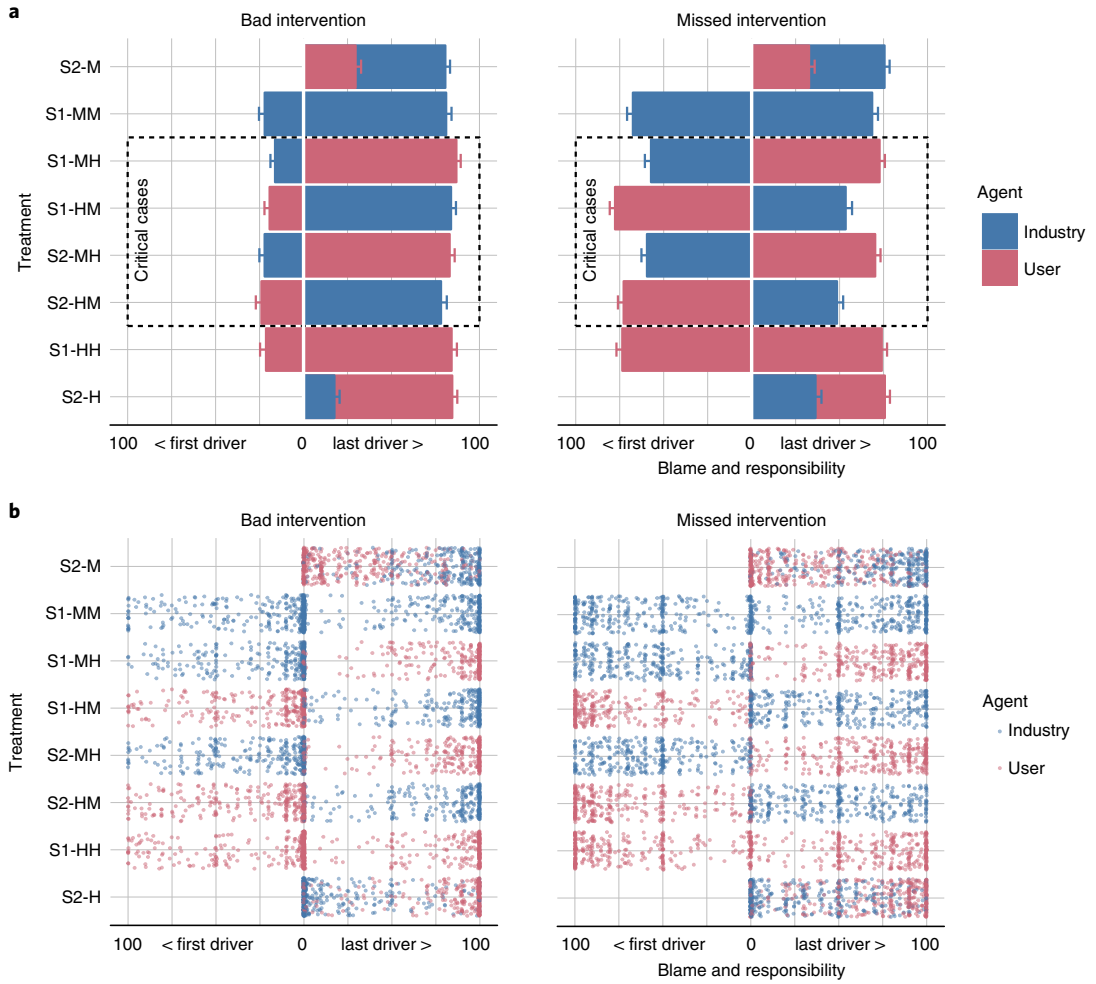

Fig. 2 | Blame ratings for user and industry in six car types. a,**b**, Bar plot (**a**) and dot plot (**b**). Data from Study 1 (S1: $n = 786$, observations = 3,144) and Study 2 (S2: $n = 382$, observations = 1,528). Ratings of blame and causal responsibility are aggregated (collectively referred to as blame, henceforth). Ratings of car and company are aggregated (collectively referred to as industry, henceforth). The *y* axis represents the six car types considered in S1 and S2. Two car types, HM (human–machine) and MH (machine–human), were considered in both studies. The *y*-axis labels include the study and the car type. For example, S1-HM represents the human–machine regime ratings collected in Study 1. In the six car types, the *x*-axis labelling of first driver refers to the main driver, while the last driver refers to the secondary driver in dual-driver cars and the sole driver in sole-driver cars. For bad intervention, only one agent has erred (the last driver). This agent (whether user or industry) is blamed more than the other agent (first driver, see Table 1). For missed intervention, in dual-driver cars (rows 2–7), both agents have erred. When human and machine are sharing control (within the dotted rectangle), blame ratings of Industry drop significantly regardless of the role of the machine (main or secondary). In Study 1, blame to Industry in S1-MH ($m_1 = 57.2$) is significantly less than in S1-MM ($m_2 = 68$), ($t(760.6) = -5.05$, $P < 0.0001$, $m_2 - m_1 = 10.8$, 95% confidence interval (CI) for $m_2 - m_1 = 6.6$–15). Blame to Industry in S1-HM ($m_1 = 53.4$) is significantly less than in S1-MM ($m_2 = 68$), ($t(722.77) = -6.6042$, $P < 0.0001$, $m_2 - m_1 = 14.6$, 95% CI for $m_2 - m_1 = 10.2$–19). In Study 2, blame to Industry in S2-M ($m_1 = 75.6$) is significantly more than in S2-MH ($m_2 = 59.5$), ($t(754.63) = -7.3885$, $P < 0.0001$, $m_1 - m_2 = 16.1$, 95% CI for $m_1 - m_2 = 11.8$–20.3) and is significantly more than in S2-HM ($m_3 = 48.51$), ($t(745.06) = 11.676$, $P < 0.0001$, $m_1 - m_3 = 27.1$, 95% CI for $m_1 - m_3 = 22.5$–31.6).

error was a significant predictor of ratings (Table 1, bad intervention, Study 2 and Fig. 2).

Missed interventions. We again replicated the main finding of Study 1. Driver type—whether the driver is a human or machine—has a significant impact on ratings. Specifically, in shared-control scenarios (machine–human and human–machine), where both human and machine have made errors, the machine driver is consistently blamed less than the human driver (Table 1, missed intervention, Study 2 and Fig. 2).

As we noted in Study 1, the human–machine difference is driven by a reduction in the blame attributed to machines when there is a human in the loop. This is verified in Study 2 by comparison of blaming the machine in the shared control cases to blaming it in the fully automated car (driven by a sole machine driver). In each case, blaming the machine in the shared control case is significantly lower than blaming the machine in the fully automated car: fully automated ($m_1 = 75.6$) versus machine–human ($m_2 = 59.5$), $t(754.63) = -7.4$, $P < 0.0001$, $m_1 - m_2 = 16.1$, 95% CI for $m_1 - m_2 = 11.8$–20.3; versus human–machine ($m_3 = 48.5$), $t(745.06) = 11.7$, $P < 0.0001$, $m_1 - m_3 = 27.1$, 95% CI for $m_1 - m_3 = 22.5$–31.6.

Study 3. In Study 3, we used the same car regimes as in Study 2 but the cases were dilemma scenarios in which the drivers had to choose between crashing into a single pedestrian or crashing into five pedestrians. This study was conducted as a comparison to Studies

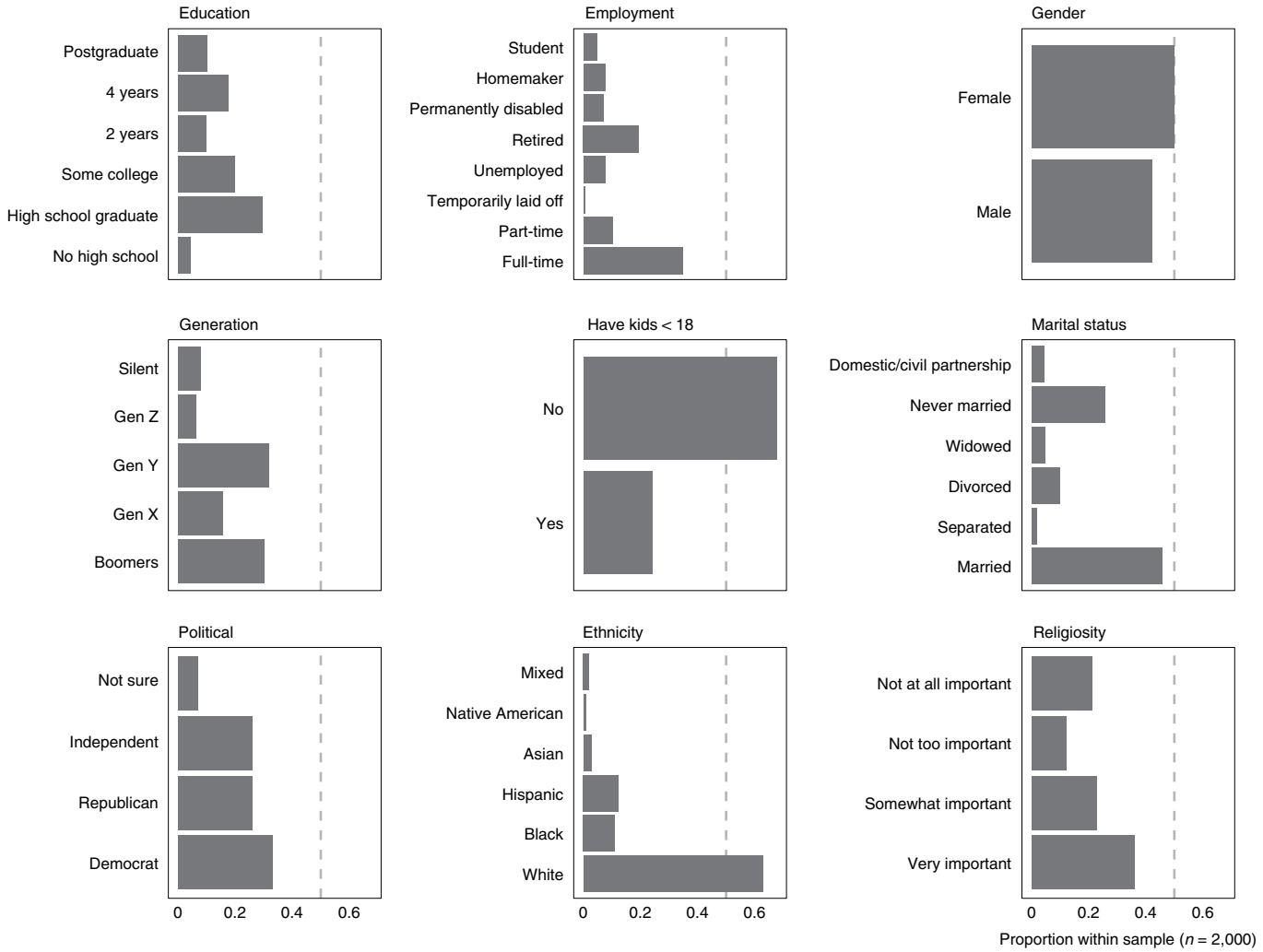

Fig. 3 | Representation of demographic attributes in Study 5. Study 5 was run via YouGov, a service that administers and runs surveys on nationally representative samples. The sample includes 2,000 participants with diverse demographic attributes. Participants who did not disclose their demographic information are not shown. Generation labels Silent, Boomers, Gen X, Gen Y, and Gen Z refer to individuals born before 1946, during 1946–1964, during 1965–1976, during 1977–1995, and after 1995, respectively.

1 and 2, which involve clear errors, and the studies conducted in previous research on self-driving cars (such as in refs. [11,16–18]), which involve the difficult choice of deciding which of two groups of people to hit. All the main effects in Study 2 were replicated in Study 3.

Bad interventions. Here, we replicated the main results of Studies 1 and 2, in bad intervention cases for the shared-control cars, whether or not the driver made an error was a significant predictor of ratings (Table 1, bad intervention, Study 3).

Missed interventions. Replicating the main results of Studies 1 and 2, driver type has a significant impact on ratings. Specifically, in shared-control scenarios where both human and machine have made errors, the machine driver is consistently blamed less than the human driver (Table 1, missed interventions, Study 3).

Study 4. In Study 4 we replicated the central findings but using more ecologically valid stimuli. We used only the human–machine and machine–human shared control cars in the missed interventions scenario; these are the cases where we observed systematic decrease in blaming the machine in Studies 1–3. For this study, we continued to use hypothetical scenarios but the stimuli shown to participants looked like realistic newspaper articles (see Supplementary methods 1, Studies 4–5).

The main finding of Studies 1–3 was replicated: the machine driver is consistently blamed less than the human driver in these shared-control scenarios where both human and machine have made errors (Table 1, missed interventions, Study 4).

Study 5. Study 5 was a replication of Study 4, run via YouGov with a nationally representative sample of the US population (see Fig. 3 for details).

The main finding was again replicated: the machine driver is consistently blamed less than the human driver (Table 1, missed interventions, Study 5). This result (that is, human is blamed more than machine) holds directionally in 82% of demographic subgroups of participants (see Fig. 4).

Discussion

Our central finding is that in cases where a human and a machine share control of the car in hypothetical scenarios, less blame is attributed to the machine when both drivers make errors. The first deadly crashes of automated vehicles (mentioned above) were similar in structure to our missed intervention cases. In those cases,

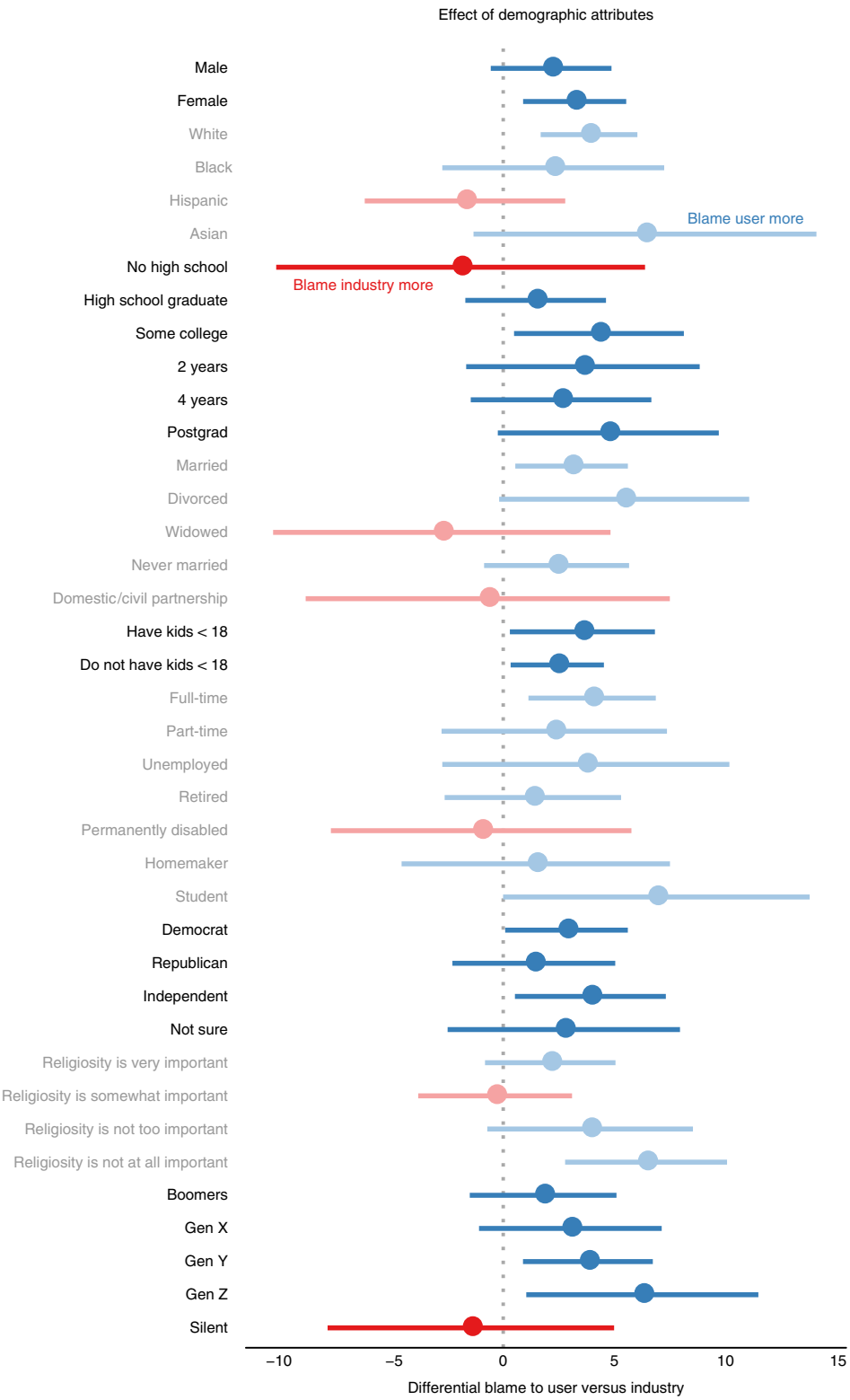

Fig. 4 | Ratings of demographic subgroups in Study 5. Data collected in Study 5 from nationally representative sample ($n = 2{,}000$, observations = 4,000). Each row represents the mean of differential blame attributed to user (that is, human) versus industry (that is, car or company). Positive values (blue) indicate more blame attributed to user while negative values (red) represent more blame attributed to industry. Error bars are 95% CI. Only subgroups with at least 50 participants are shown; 33 of 40 subgroups (82%) attribute more blame to user. Few subgroups blame industry more, and those that do typically have smaller samples.

both the machine primary driver and human secondary driver should have taken action to avoid a collision, and neither driver did. Our results suggests that the public response that occurred to the crash—one that focused attention on the driver being exceedingly negligent—is likely to generalize to other dual-error missed intervention-style cases, shifting blame away from the machine

and towards the human. Moreover, the convergence of our results with this real-world public reaction seems to suggest that while we employed stylized, simplified vignettes in our research, our findings show external validity. Moreover, this pattern of results was replicated in a nationally representative sample of the US population (and across different subgroups; see Fig. 4), which employed naturalistic presentation of scenarios (see Supplementary methods 1, Studies 4–5).

Our central finding (diminished blame apportioned 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. Because the public are less likely to see the machine as being at fault in dual-error cases like the Tesla and Uber crashes, the sort of public pressure that drives regulation might be lacking. For instance, if we were to allow the standards for automated vehicles to be set through jury-based court-room decisions, we expect that juries will be biased to absolve the car manufacturer of blame in dual-error cases, thereby failing to put sufficient pressure on manufacturers to improve car designs. Despite the fact that there are some avenues available to courts to mitigate psychological biases that may arise among juries (such as carefully worded jury instructions or expert witnesses), psychological biases continue to play an important role in court-based decisions[20]. In fact, we have been in a similar situation before. Before the 1960s, car manufacturers enjoyed a large amount of liberty from liability when a car's occupant was harmed in a crash (because blame in car crashes was attributed to the driver's error or negligence). Top-down regulation was necessary to introduce the concept of 'crash worthiness' into the legal system—that is, that cars should be designed in such a way as to minimize injury to occupants when a crash occurs. Only following these laws were car manufacturers forced to improve their designs[21]. Here, too, top-down regulation of automated car safety might be needed to correct a public under-reaction to crashes in shared-control cases. What, exactly, the safety standard should be is still an open question, however.

If our data identify a source of possible public over-reaction, it is for cars with a human primary driver and a machine secondary driver in bad intervention-style cases. These are the only cases we identified where the car receives more blame than the human. It seems possible that these sorts of cars may generate widespread public concern once we see instances of bad intervention-style crashes in human–machine car regimes. This could potentially slow the transition to fully automated vehicles if this reaction is not anticipated and managed appropriately in public discourse and legal regulation. Moreover, manufacturers that are working to release cars with a machine secondary driver should plan appropriately for the probable legal fall-out for these unique cases where that driver receives more blame than a human.

Our data portend the sort of reaction we can expect to automated car crashes at the societal level (for example, through public reaction and pressure to regulate). Once we begin to see societal-level responses to automated cars, that reaction may shape incentives for individual actors. For example, people may want to opt into systems that are designed such that, in the event of a crash, the majority public response will be to blame the machine. Worse yet, people may train themselves to drive in a way that, if they crash, the blame is likely to fall to the machine (for instance, by not attempting to correct a mistake that is made by a machine override). This sort of incentive shaping may already be happening in the legal domain. Judges who make decisions about whether to release a person from custody between arrest and trial frequently rely on actuarial risk assessment tables to help make their decision. Some suspect that judges may be overly reliant on these tables as a way of diminishing their responsibility if a released person commits a crime. Recent attention generated in response to such a case focused on the role of the algorithm rather than the judge[22], indicating that the possibility of incentive shaping in the legal domain is not so far-fetched.

Given these possible societal-level implications of our findings, it is important to acknowledge the potential limitations of interpreting our data this broadly. First, the participants in all of our experiments know that they are reading about hypothetical scenarios. It is possible that this reduces the psychological realism of the study[23,24], causing participants' responses to be characteristically different to what they would be after reading about an actual event. The literature provides a mixed view of how well responses to hypothetical scenarios map onto those made in real-life situations[25–28]. However, the research that does show considerable differences[25,26] finds that these differences are mostly seen in the way participants themselves would act in moral situations and not necessarily about the moral judgments they render about third parties. In our paper, we study participants' judgements (blame and causal responsibility) about third parties in hypothetical scenarios; these may align more directly with judgements of actual scenarios.

Second, although we may see a reasonably tight mapping between the opinions expressed in this study's scenarios and those that would be expressed in real-life situations, it is important to note that, in the latter case, judgements will not be occurring in isolation. Instead, they will occur within a richer context than the carefully controlled scenarios used in our studies. People may hear reports of accidents with more emotion-arousing details, which are known to skew people's judgements[29,30]. Moreover, the public's reaction to hearing about semi-autonomous vehicle crashes will be shaped by many factors beyond their immediate psychological response (which is the object of our study), including opinion pieces they read, the views of community leaders and so on. These factors will collectively shape the public's overall reaction to crashes.

Studies 1, 2, 4 and 5 looked at blame and causal responsibility attribution in cases where one or both drivers made errors. Study 3 looked at dilemma scenarios where the drivers faced the choice of running over either one or five pedestrians. While there is, in some sense, an 'optimal' outcome in these cases (corresponding to saving more lives), it is not obvious that it would (for example) count as an error to refuse to swerve away from five pedestrians into a pedestrian that was previously unthreatened. In fact, the German Ethics Commission on Automated and Connected Driving report[31] indicates that programming cars to trade off lives in this way would be prohibited. The report states: "It is also prohibited to offset victims against one another. (…) Those parties involved in the generation of mobility risks must not sacrifice non-involved parties". Even though participants in previous studies prefer to sacrifice one person who was previously not involved than five (for example, refs. [16,17]), the German Ethics Commission's decision underscores the fact that trading off lives in dilemma situations can be particularly fraught. For this reason, and for continuity with previous work on the ethics of self-driving cars[11,16,17] and in moral psychology more generally[32,33], we chose to investigate dilemma situations. Our findings about the effect of driver type in these cases underscore the fact that findings about how blame and responsibility are attributed after a crash may still hold in less-clear dilemma scenarios.

Some of our results fall in line with previous work on the psychology of causal inference. In bad intervention cases, the primary driver (be it human or machine) makes a correct decision to keep the car on a course that will avoid a pedestrian. Following this, the secondary driver makes the decision to swerve the car into the pedestrian. Our data show that the secondary driver (the one that makes a mistake) is considered more causally responsible than the first. It is well established that judgements of causal responsibility are impacted by violations of statistical and moral norms[34–37], and a mistake seems to count as such a violation. That is, if something unusual or counter-normative happens, that event is more likely to

be seen as a cause of some effect than another event that is typical or norm-conforming.

Moreover, the central finding that humans are blamed more than machines, even when both make errors, accords with research on the psychology of causal attribution. Findings in that field suggest that voluntary causes (causes created by agents) are better causal explanations than physical causes[38]. While it is clear that what a human does is fundamentally different to what a machine does in each of the scenarios, it remains an open question whether an artificial intelligence that is operating a car is perceived as a physical cause, an agent, something in between or something else entirely[39,40]. Future work should investigate the mental properties attributed to an artificial intelligence that controls a car both in conjunction with a human and alone. Understanding the sort of mind we perceive as dwelling inside an artificial intelligence may help us understand and predict how blame and causal responsibility will be attributed to it[41].

Another open question concerns the implications of attributing blame to a machine at all. There are various ways that humans express moral condemnation. For example, we may call an action morally wrong, say that a moral agent has a bad character or judge that an agent is blameworthy. Judgements of blame typically track judgements of willingness to punish the perpetrator[42,43]. Are the participants in our study expressing that some punishment is due to the machine driver of the car, whatever that may mean? Alternately, is it possible that participants' expressions of blame indicate that some entity is deserving of punishment that represents the machine (the company, or a human representative of the company, such as the chief executive officer). The similar blame judgements given to the car and the car's representatives (company) perhaps support this possibility. Finally, it is possible that participants ascribe only non-moral blame to the machine, in the sense of it being responsible but not in a moral sense. We may say that a forest fire is to blame for displacing residents from their homes, without implying that punishment is due to anyone at all.

Following these studies, the reason that participants blame machine drivers less than human drivers in missed intervention cases also remains an open question. The findings may be linked to the uncertainty with which we perceive the agential status of machines. Once machines are a more common element in our moral world and we interact with them as moral actors, will this effect change? Or will this finding be a lasting hallmark of the cognitive psychology of human–machine interaction?

A final open question concerns whether the effects we report here will generalize to other cases of human–machine interaction. Already we see fruitful human–machine partnerships emerging with judges, doctors, military personnel, factory workers, artists and financial analysts, to name but a few. We conjecture that we may see the patterns we report here in domains other than automated vehicles, though each domain will have its own complications and quirks as machines begin to become more subtly integrated in our personal and professional lives.

Methods

This study was approved by the Institute Review Board at Massachusetts Institute of Technology. The authors complied with all relevant ethical considerations, including obtaining informed consent from all participants.

In all studies, participants were allocated uniformly randomly into conditions. Data collection and analysis were performed blind to the conditions of the experiments. The sample size was chosen in each study to ensure the inclusion at least 100 participants for each condition. Numbers of participants were chosen in advance of running the study, and all data were collected before analysis. See details below.

In Studies 1–3 we excluded any participant who did not (1) complete all measures within the survey, (2) transcribe (near-perfectly) a 169-character paragraph from an image (used as an attention check) and (3) have a unique MTurk ID per study (all records with a recurring MTurk ID were excluded).

Case description. Summary descriptions of all car types and cases. For full vignettes, see Supplementary methods 1.

Sole-driver car. This car has only one driver that does all the driving. Two versions are used.

Human-only. This is a sole-driver car, in which a human is the driver. Also referred to as a regular car.

Machine-only. This is a sole-driver car, in which a machine is the driver. Also referred to as a fully automated car.

Dual-driver car. This car has a primary driver whose job it is to drive the car, and a secondary driver whose job it is to monitor the actions of the first driver and intervene when the first driver makes an error (also referred to as shared-control car). Four versions are used.

Human–machine. This is a dual-driver car in which a human is the primary driver and a machine is the secondary driver (also referred to as Guardian).

Machine–human. This is a dual-driver car in which a machine is the primary driver and a human is the secondary driver (also referred to as Autopilot).

Human–human. This is a dual-driver car in which a human is the primary driver and another human is the secondary driver.

Machine–machine. This is a dual-driver car in which a machine is the primary driver and another machine is the secondary driver.

Intervention types. We use two types of intervention: bad intervention and missed intervention. The description of each is dependent on whether the car is a sole- or a dual-driver car.

Bad intervention (dual-driver). The primary driver kept the car on its track. The secondary driver intervened and steered the car off its track (killing a pedestrian) rather than keeping the car on track and killing no one.

Missed intervention (dual-driver). The primary driver kept the car on its track. The secondary driver kept the car on its track (killing a pedestrian) rather than swerving into the adjacent lane and killing no one.

Bad intervention (sole-driver). The sole driver steered the car off its track (killing a pedestrian) rather than keeping the car on track and killing no one.

Missed intervention (sole-driver). The sole driver kept the car on its track (killing a pedestrian) rather than swerving into the adjacent lane and killing no one.

Dilemma versions (Study 3). The two outcomes of killing one pedestrian versus killing no one are replaced with the two outcomes of killing five pedestrians versus killing one pedestrian. For example, in missed intervention (dual-driver): (…) The secondary driver kept the car on its track (killing five pedestrians) rather than swerving into the adjacent lane and killing one pedestrian.

Study 1. *Participants.* The data were collected in September 2017 from 809 participants (US residents) recruited from the Mechanical Turk platform (each was compensated US$0.5). Of those, 23 participants were excluded (as explained above) leaving us with 786 participants. Participants were aged 18–83 years (median, 33 years), 50% were females, 39% had an annual income of US$50,000 or more and 55% had a bachelor degree or higher.

Stimuli and procedures. Participants were uniformly randomly allocated to one of four conditions. Conditions varied the car type (human–human, human–machine, machine–human and machine–machine) in a four-level between-subjects design. In each condition, participants first read a description of the car and were then asked to attribute competence to each of the two drivers on a 100-point scale anchored at 'not competent' and 'very competent' (see Supplementary Fig. 1 for results on competence). Participants then read two scenarios (presented in a random order), one bad intervention case and one missed intervention case. After each scenario, participants were asked to indicate (on a 100-point scale) to what extent they thought each driver was blameworthy (from 'not blameworthy' to 'very blameworthy'), and to what degree each of these two agents caused the death of the pedestrian (from 'very little' to 'very much'). Questions were presented in a randomized order. (See Supplementary methods 1, Study 1 for text of the vignettes and see Supplementary methods 2 for questions). At the end of the surveys, participants provided basic demographic information (for example, age, gender, income, education).

Study 2. *Participants.* The data were collected in May 2017 from 804 participants (US residents) recruited from the Mechanical Turk platform (each was compensated US$0.3). Of those, 25 participants were excluded (as explained above), leaving us with 779 participants. Participants were aged 18–77 years (median, 32 years), 48% were females, 39% had an annual income of US$50,000 or more and 54% had a bachelor degree or higher.

Stimuli and procedures. Participants were uniformly randomly allocated to one of eight conditions. Conditions varied the car type (human-only, human–machine, machine–human and machine only) and the industry representative (car and company), in a 4×2 between-subjects multifactorial design. In each condition, participants read two scenarios (presented in a random order), one bad intervention case and one missed intervention case. After each scenario, participants were asked to attribute causal responsibility, blameworthiness and competence (see Supplementary Fig. 1 for results on competence) to two agents: the human in the car and a representative of the car (the car itself or the manufacturing company of the car, depending on the condition). All other features of Study 2 were the same as those in Study 1.

Study 3. *Participants.* The data were collected in November 2016 from 1,008 participants (US residents only) recruited from the Mechanical Turk platform (each was compensated US$0.6). Of those, 35 participants were excluded (as explained above), leaving us with 973 participants. Participants were aged 18–84 years (median, 33 years), 51% were females, 37% had an annual income of US$50,000 or more and 53% had a bachelor degree or higher.

Stimuli and procedures. There were two groups of participants in Study 3: those who saw dual-driver cases and those who saw sole-driver. For those who saw dual-driver cases, participants were randomly assigned to one of six conditions in a 2×3 design, varying the car type (human–machine or machine–human) and the industry representative (car, company and programmer). Data for the programmer were later dropped from the analysis. For those who saw sole-driver cases, participants were randomly assigned to one of four conditions in a 2×2 design, varying the car type (human-only or machine-only) and the industry representative (car or company). In each condition (for both dual- and single-car groups), participants read two scenarios (presented in a random order), one bad intervention case and one missed intervention case. These scenarios were the dilemma versions of those presented in Studies 1 and 2 (see description above). After each scenario, participants were asked to attribute causal responsibility and blameworthiness to two agents: the human in the car and a representative of the car (the car itself, the company or the programmer, depending on the condition). All other features of Study 3 were identical to those of Study 2.

Study 4. *Participants.* The data were collected in January 2019 from 375 participants (US residents only) recruited from the Mechanical Turk platform (each was compensated US$0.3). No demographic data were collected for this study. Given that it was done on the same platform as Studies 1–3 (that is, Mechanical Turk), its demographic proportions are expected to be similar.

Stimuli and procedures. The key elements of this study and Study 5 are (1) the restriction to missed intervention cases and (2) the visual and textual content of the vignettes have the look and feel of a news piece (see Supplementary methods 1, Studies 4–5).

Participants were uniformly randomly allocated to one of four conditions. Conditions varied the car type (human–machine and machine–human) and the industry representative (car and company), in a 2×2 between-subjects multifactorial design. In each condition, participants read one scenario—one missed intervention case. The textual content of these scenarios was close to that presented in Studies 1–3, with slight changes to make it read like a news piece. After each scenario, participants were asked to attribute blameworthiness to two agents: the human in the car and a representative of the car (the car itself or the company, depending on the condition).

Study 5. *Participants.* The data were collected in March 2019 from 2,189 participants (US residents) recruited via YouGov, a service that administered the study and collected the data from a representative sample of participants. The participants were then matched down to a sample of 2,000 participants based on demographics. See Fig. 3 for details on demographic proportions of participants in this study.

Stimuli and procedures. This study is identical in set-up in Study 4.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw data and source data for Figs. 2–4, Table 1 and Supplementary Fig. 1 are available at https://bit.ly/2kzLymH.

Code availability

Code used to produce figures and tables in this article is available at https://bit.ly/2kzLymH.

Received: 23 March 2018; Accepted: 24 September 2019;
Published online: 28 October 2019

References

1. Road traffic injuries. *World Health Organization Fact Sheet* (WHO, 2017).
2. Geistfeld, M. A. A roadmap for autonomous vehicles: state tort liability, automobile insurance, and federal safety regulation. *Calif. L. Rev.* **105**, 1611 (2017).
3. Tesla. *A Tragic Loss* https://www.tesla.com/blog/tragic-loss (Tesla, 2016).
4. *Automatic Vehicle Control Systems—Investigation of Tesla Accident* (National Highway Traffic Safety Administration, 2016).
5. Griswold, A. Uber found not criminally liable in last year's self-driving car death. *Quartz* (5 March 2019).
6. Lowy, J. & Krishner, T. Tesla driver killed while using autopilot was watching Harry Potter, witness says. *Associated Press News* https://apnews.com/ee71bd075fb948308727b4bbff7b3ad8 (30 June 2016).
7. Chong, Z. & Krok, A. Tesla not at fault in fatal crash, driver was not watching a movie. *CNET* https://www.cnet.com/roadshow/news/tesla-found-not-guilty-of-fatal-crash-says-us-report/ (19 June 2017).
8. Randazzo, R. Who was really at fault in fatal uber crash? here's the whole story. *AZ Central* https://www.azcentral.com/story/news/local/tempe/2019/03/17/one-year-after-self-driving-uber-rafaela-vasquez-behind-wheel-crash-death-elaine-herzberg-tempe/1296676002/ (17 March 2019).
9. Munster, G. Here's when having a self-driving car will be a normal thing. *Fortune* https://fortune.com/2017/09/13/gm-cruise-self-driving-driverless-autonomous-cars/ (13 September 2017).
10. Kessler, S. A timeline of when self-driving cars will be on the road, according to the people making them. *Quartz* https://qz.com/943899/a-timeline-of-when-self-driving-cars-will-be-on-the-road-according-to-the-people-making-them/ (29 March 2017).
11. Li, J., Zhao, X., Cho, M.-J., Ju, W. & Malle, B. F. *From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-driving Cars* SAE Technical Paper 2016-01-0164 https://doi.org/10.4271/2016-01-0164 (SAE, 2016).
12. Chockler, H. & Halpern, J. Y. Responsibility and blame: a structural-model approach. *J. Artif. Intell. Res.* **22**, 93–115 (2004).
13. Gerstenberg, T. & Lagnado, D. A. When contributions make a difference: explaining order effects in responsibility attribution. *Psychon. Bull. Rev.* **19**, 729–736 (2012).
14. Sloman, S. A. & Lagnado, D. Causality in thought. *Annu. Rev. Psychol.* **66**, 223–247 (2015).
15. Zultan, R., Gerstenberg, T. & Lagnado, D. A. Finding fault: causality and counterfactuals in group attributions. *Cognition* **125**, 429–440 (2012).
16. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
17. Awad, E. et al. The moral machine experiment. *Nature* **563**, 59 (2018).
18. Malle, B., Scheutz, M., Arnold, T., Voiklis, J. & Cusimano, C. Sacrifice one for the good of many? People apply different. In *Proc. 10th ACM/IEEE International Conference on Human-Robot Interaction* 117–124 (2015).
19. Shariff, A., Bonnefon, J.-F. & Rahwan, I. Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Hum. Behav.* **1**, 694 (2017).
20. Bornstein, B. H. & Greene, E. Jury decision making: implications for and from psychology. *Curr. Dir. Psychol. Sci.* **20**, 63–67 (2011).
21. Nader, R. *Unsafe at Any Speed. The Designed-in Dangers of the American Automobile* (Grossman, 1965).
22. Westervelt, E. Did a bail reform algorithm contribute to this San Francisco man's murder? *National Public Radio* https://www.npr.org/2017/08/18/543976003/did-a-bail-reform-algorithm-contribute-to-this-san-francisco-man-s-murder (18 August 2017).
23. Bauman, C. W., McGraw, A. P., Bartels, D. M. & Warren, C. Revisiting external validity: concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Soc. Personal Psychol. Compass* **8**, 536–554 (2014).
24. Aronson, E, Wilson, T. D. & Brewer, M. B. in *The Handbook of Social Psychology* Vol. 1 (eds Gilbert, D. T., Fiske, S. T., & Lindzey, G.) 99–142 (McGraw-Hill, 1998).
25. FeldmanHall, O. et al. Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Soc. Cogn. Affect. Neurosci.* **7**, 743–751 (2012).
26. Bostyn, D. H., Sevenhant, S. & Roets, A. Of mice, men, and trolleys: hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychol. Sci.* **29**, 1084–1093 (2018).
27. Dickinson, D. L. & Masclet, D. Using ethical dilemmas to predict antisocial choices with real payoff consequences: an experimental study. *J. Econ. Behav. Organ.* **166**, 195–215 (2018).
28. Plunkett, D. & Greene, J. Overlooked evidence and a misunderstanding of what trolley dilemmas do best: a comment on Bostyn, Sevenhant, & Roets (2018). *Psychol. Sci.* **30**, 1389–1391 (2019).
29. Greene, J. & Haidt, J. How (and where) does moral judgment work? *Trends Cogn. Sci.* **6**, 517–523 (2002).
30. Horberg, E. J., Oveis, C. & Keltner, D. Emotions as moral amplifiers: an appraisal tendency approach to the influences of distinct emotions upon moral judgment. *Emot. Rev.* **3**, 237–244 (2011).

31. Luetge, C. The German ethics code for automated and connected driving. *Philos. Technol.* **30**, 547–558 (2017).
32. Mikhail, J. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment* (Cambridge Univ. Press, 2011).
33. Greene, J. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them* (Penguin, 2014).
34. Alicke, M. D. Culpable control and the psychology of blame. *Psychol. Bull.* **126**, 556 (2000).
35. Gerstenberg, T., Goodman, N. D., Lagnado, D. A. & Tenenbaum, J. B. How, whether, why: causal judgments as counterfactual contrasts. in *Proc. 37th Annual Meeting of the Cognitive Science Society* 782–787 (2015).
36. Hitchcock, C. & Knobe, J. Cause and norm. *J. Philos.* **106**, 587–612 (2009).
37. Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. & Knobe, J. Causal superseding. *Cognition* **137**, 196–209 (2015).
38. Hart, H. L. A. & Honoré, T. *Causation in the Law* (Oxford Univ. Press, 1985).
39. Gray, H. M., Gray, K. & Wegner, D. M. Dimensions of mind perception. *Science* **315**, 619–619 (2007).
40. Weisman, K., Dweck, C. S. & Markman, E. M. Rethinking people's conceptions of mental life. *Proc. Natl Acad. Sci. USA* **114**, 11374–11379 (2017).
41. Gray, K., Young, L. & Waytz, A. Mind perception is the essence of morality. *Psychol. Inq.* **23**, 101–124 (2012).
42. Cushman, F. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* **108**, 353–380 (2008).
43. Cushman, F. Deconstructing intent to reconstruct morality. *Curr. Opin. Psychol.* **6**, 97–103 (2015).

Acknowledgements

I.R., E.A., S.L. and S.D. acknowledge support from the Ethics and Governance of Artificial Intelligence Fund. J.-F.B. acknowledges support from the ANR-Labex Institute for Advanced Study in Toulouse, the ANR-3IA Artificial and Natural Intelligence Toulouse Institute and grant no. ANR-17-EURE-0010 from Investissements d'Avenir. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

E.A., S.L., M.K.-W., S.D., J.B.T., A.S., J.-F.B. and I.R. contributed to the conception and design of the research. E.A., S.L., M.K.-W. and S.D. conducted studies. E.A. and J.-F.B. analysed data. S.L., E.A., M.K.-W., J.-F.B. and I.R. wrote the manuscript. All authors reviewed and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0762-8.

Correspondence and requests for materials should be addressed to J.B.T., A.S., J.-F.B. or I.R.

Peer review information Primary Handling Editor: Mary Elizabeth Sutherland

nature research

Corresponding author(s): Edmond Awad

Last updated by author(s): Sep 9, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For studies 1-4, data was collected through Qualtrics, and participants were recruited from Amazon Mechanical Turk. For study 5, data was collected and administered by YouGov, a service that recruits nationally representative samples.
Data analysis	Data was analyzed using R (RStudio 3.4.1)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw data and Source data for Fig 2, 3 and 4; Table 1; and Supplementary Fig 1 are available in: https://bit.ly/2kzLymH

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The submitted manuscript present results from five studies. Are all quantitative studies.
Research sample	For Studies 1-4, participants are recruited through Amazon Mechanical Turk (AMT), an online platform. This platform allows for the collection of high-quality data from a considerably large sample with minimal cost and effort. This platform is widely used to collect data in psychology and social surveys. As for demographics: Study 1 has 786 participants. Participants are aged between 18-83 (median: 33), 50% are females, 39% has annual income of $50K or more, and 55% has a bachelor degree or higher. Study 2 has 779 participants. Participants are aged between 18-77 (median: 32), 48% are females, 39% has annual income of $50K or more, and 54% has a bachelor degree or higher. Study 3 has 931 participants. Participants are aged between 18-84 (median: 33), 51% are females, 37% has annual income of $50K or more, and 53% has a bachelor degree or higher. Study 4 has 375 participants. No demographic data was collected for this study. Given that it was done on the same platform as studies 1-3 (i.e. Mechanical Turk), its demographic proportions are expected to be similar. For Study 5, participants are recruited by YouGov, a service that runs and administers experiments using nationally representative samples. Study 5 has 2000 participants. Participants are aged between 19-89 (median: 49), 53% are females, 48% are currently employed (full or part time), and 30% has a bachelor degree or higher. The demographic proportion for this study is available in Supplemental Figure 4.
Sampling strategy	For studies 1-4, like other studies usually run on AMT, the survey was advertised on the platform and was available on a first come, first served basis. As a rule of thumb, the sample size was chosen in each study as to ensure having 100 participants for each condition. In one study, 200 participants were considered for each condition. For study 5, the sample size was 500 participants for each of the four conditions.
Data collection	Studies 1-4 were programmed on Qualtrics survey software and participants (USA residents only) were recruited from the online platform AMT. Allocation to conditions was handled by the software. Participants who did one study were not allowed to enter any of the following studies. Study 5 was run and administered by YouGov (https://yougov.co.uk/solutions/research).
Timing	Study 1: September 2017, Study 2: May 2017, Study 3: November 2016, Study 4: January 2019, Study 5: March 2019
Data exclusions	For studies 1-3, we excluded any subjects who did not (i) transcribe (near-perfectly) a 169-character paragraph from an image (used to exclude non-serious Turkers), and (ii) have unique TurkID per study (all records with a recurring MTurk ID were excluded). Study 1 excluded data of 23 participants, Study 2 excluded data of 25 participants, and Study 3 excluded data of 35 participants. For study 5, YouGov interviewed 2189 respondents who were then matched down to a sample of 2000 to produce the final dataset. The respondents were matched to a sampling frame on gender, age, race, and education. The frame was constructed by stratified sampling from the full 2016 American Community Survey (ACS) 1-year sample with selection within strata by weighted sampling with replacements (using the person weights on the public use file). The matched cases were weighted to the sampling frame using propensity scores. The matched cases and the frame were combined and a logistic regression was estimated for inclusion in the frame. The propensity score function included age, gender, race/ethnicity, years of education, and region. The propensity scores were grouped into deciles of the estimated propensity score in the frame and post-stratified according to these deciles.
Non-participation	380 participants (328 before and 52 after condition allocation) dropped out from Study 1; 347 participants (276 before and 71 after condition allocation) dropped out from Study 2; 327 participants (220 before and 107 after condition allocation) dropped out from Study 3; and 68 participants (37 before and 31 after condition allocation) dropped out from Study 4. Dropouts are participants who dropped after seeing the very first page of the survey, but before finishing the last question in the survey. All data of dropouts were excluded from studies.
Randomization	In all studies, participants were allocated uniformly randomly into conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	For studies 1-4, participants were recruited from Amazon Mechanical Turk. The recruited sample is not nationally representative, but given that this platform is widely used to collect data in psychology and social surveys, the sample is comparable to many currently available studies. For study 5, recruitment was done by YouGov, a service that recruits nationally representative samples.
Ethics oversight	This study was approved by the Institute Review Board (IRB) at Massachusetts Institute of Technology (MIT). The authors complied with all relevant ethical considerations.

Note that full information on the approval of the study protocol must also be provided in the manuscript.