Risk Analysis, Vol. 00, No. 0, 2021 DOI: 10.1111/risa.13674

Blame Attribution Asymmetry in Human–Automation Cooperation

Peng Liu* and Yong Du

Human–automation cooperation has become ubiquitous. In this concept, automation refers to autonomous machines, robots, artificial intelligence, and other autonomous nonhuman agents. A human driver will share control of semiautonomous vehicles (semi-AVs) with an automated system and thus share responsibility for crashes caused by semi-AVs. Research has not clarified whether and why people would attribute different levels of blame and responsibility to automation (and its creators) and its human counterpart when each causes an equivalent crash. We conducted four experiments in two studies (total $N = 1{,}045$) to measure different responses (e.g., severity and acceptability judgment, blame and responsibility attribution, compensation judgment) to hypothetical crashes that are caused by the human or the automation in semi-AVs. The results provided previously unidentified evidence of a bias, which we called the "blame attribution asymmetry," a tendency that people will judge the automation-caused crash more harshly, ascribe more blame and responsibility to automation and its creators, and think the victim in this crash should be compensated more. This asymmetry arises in part because of the higher negative affect triggered by the automation-caused crash. This bias has a direct policy implication: a policy allowing "not-safe enough" semi-AVs on roads could backfire, because these AVs will lead to many traffic crashes, which might in turn produce greater psychological costs and deter more people from adopting them. Other theoretical and policy implications of our findings were also discussed.

KEY WORDS: affect heuristic; blame and responsibility attribution; blame attribution asymmetry; human–automation cooperation; semiautonomous vehicles

1. INTRODUCTION

Human–automation cooperation has become ubiquitous in people's daily lives and working environments (here, automation, autonomous machines and robots, and artificial intelligence [AI] are interchangeable) (Bigman, Waytz, Alterovitz, & Gray, 2019). Doctors work with medical robots to seek and remove tumors. Human pilots cooperate with military drones to select a target and launch a missile strike. The present study focused on human–automation cooperation in semiautonomous vehicles (semi-AVs). A human driver and automated driving system share control of a semi-AV, in which the human can hand over the steering or acceleration functions to automation but still prepare to take control of the vehicle (Society of Automotive Engineers, 2018). Although semi-AVs promise to increase road safety (National Highway Traffic Safety Administration, 2016), current semi-AVs being tested on roads are deemed riskier than conventional human-driven vehicles (Banerjee, Jha, Cyriac, Kalbarczyk, & Iyer, 2018; Favarò, Nader, Eurich, Tripp, & Varadaraju, 2017). On March 18, 2018, a woman pedestrian was killed after being struck by an Uber's AV in Arizona(Associated Press, 2018). At present, semi-AVs

College of Management and Economics, Tianjin University, Tianjin, China.
*Address correspondence to Peng Liu, College of Management and Economics, Tianjin University, Tianjin 300072, China; tel: (86)022-27403423; fax: (86)022-27401779; pengliu@tju.edu.cn.

may be expected to lead to more traffic crashes as more of them are being tested and running on roads.

As accidents caused by semi-AVs are unavoidable, ethical issues (i.e., causality, responsibility, and blame attributions) require serious consideration (Hevelke & Nida-Rümelin, 2015). Who is responsible for a crash when it happens? Is it the AV user, AV designer and manufacturer, or the AV? These issues in cases of the highest automation (i.e., self-driving vehicles, SDVs) or low automation (e.g., assisted automation) are not as complex as those in cases of semi-AVs. Certain car manufacturers, such as Volvo (Atiyeh, 2015) and Audi (Maric, 2017), have promised that when their SDVs (i.e., fully automated vehicles) are ready to go on the market, they will take responsibility if any crashes occur. In cases of assisted automation, Tesla has claimed that the responsibility remains with the human driver after a man riding in a Tesla Model S in the "Autopilot" mode crashed into a white truck that neither the driver nor Autopilot had detected (Tesla, 2016). In the case of Uber's 2018 fatal crash, the safety driver in its AV has been charged with negligent homicide for her failure to monitor the road, whereas Uber did not face criminal liability over this tragedy (Conger, 2020), although its automated driving system detected the pedestrian several times before impact and failed to classify her as a pedestrian (National Transportation Safety Board, 2019).

A debate has arisen on blame and responsibility attribution when semi-AVs and SDVs cause crashes in the legal, ethical, and philosophical literature. Researchers have focused on the responsibility of the AV user (e.g., is the user responsible for a crash if they did not drive the car manually?) and moral agency and moral reasoning of the car as an artificial agent (Coeckelbergh, 2016). Rahwan et al. (2019) used an example to claim that machines should not bear moral responsibility for their actions: "If a dog bites someone, the dog's owner is held responsible" (p. 483). However, more researchers held that attributions of blame and responsibility depend on contextual factors (Hevelke & Nida-Rümelin, 2015). For instance, Nyholm (2018) argued that in cases where a human is supervising the driving and would take over control of the vehicle as needed, the human should be regarded as the responsible party for unexpected outcomes when traveling. In cases where the performance of an AV is monitored by its designers and makers, as Nyholm also argued, the people who make and update the AV, rather than the human traveling in the AV, are the main loci of responsibility for how the car performs.

Although the topic of blame and responsibility attribution in AVs is gaining attention, much remains unknown about them from the layperson's perspective in the context of human–automation cooperation. In the eyes of lay people, autonomous agents can be blamed or held responsible for unexpected outcomes (Malle, Thapa Magar, & Scheutz, 2019). As the final consumers of AVs, their opinions and preferences in these issues should be taken seriously (Motamedi, Wang, Zhang, & Chan, 2020). Given the growing presence of semi-AVs on public roads, the study of how people judge traffic accidents caused by semi-AVs and subsequently attribute blame and responsibility is of great societal importance. Our research aimed to elucidate how people judge traffic accidents caused by human and automation that result in equivalent consequences in semi-AVs and whether and why people ascribe different blame and responsibility to them.

2. LITERATURE REVIEW AND HYPOTHESIS DEVELOPMENT

2.1. Humans versus Autonomous Agents

People have a natural propensity to mindlessly apply social rules and expectations to autonomous agents and therefore show similar social reactions in their interaction with autonomous agents as well as other persons (Nass & Moon, 2000). Although people are averse to autonomous agents making moral decisions, whether in the military, law, driving, or medical settings (Bigman & Gray, 2018; Gogoll & Uhl, 2018), a majority of people would consider autonomous agents morally responsible for unexpected outcomes because of their moral decisions (Kahn et al., 2012; Malle et al., 2019; Shank, DeSanti, & Maninger, 2019). Judgments of moral responsibility hinge on autonomy (Bigman et al., 2019; Kahn et al., 2012; Kim & Hinds, 2006) and mind perception (e.g., ascribing mental abilities to think and feel to automation) (Bigman & Gray, 2018; Bigman et al., 2019).

Moral dilemma research relying on vignette-based design has reported differences in the ways that people treat human and autonomous agents. Facing the moral dilemma in which a human or robot must take action or remain inactive while determining the direction of a trolley, participants tend to have a higher acceptance rating when the robot takes a

utilitarian action (i.e., sacrificing one for the good of many) and assign more blame to the robot when it refrains from that decision compared with the human (Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015; Voiklis, Kim, Cusimano, & Malle, 2016). In the military context of launching a missile strike on a terrorist compound but risking the life of a child or canceling the strike to protect the child but risking a terrorist attack, a human pilot receives more blame for canceling than for launching, whereas an AI agent and autonomous drone receive roughly the same amount of blame for canceling than for launching (Malle et al., 2019). A replication of Malle et al. (2015) with Japanese student participants, however, showed participants applying the same moral norm to human and robot agents (Komatsu, 2016), probably suggesting that cultural or demographic factors play a role in moral judgments. Shank et al. (2019) reported that when human and AI agents make violations, their differences in moral permissibility is nonsignificant, but humans are morally faulted more than AI. In summary, although people would regard autonomous agents as moral agents, they could treat autonomous agents and humans differently in ethical and moral judgment and blame and responsibility attribution.

2.2. Blame and Responsibility Attribution in AV Crashes

Certain vignette-based studies (Awad et al., 2020; Kohn, Quinn, Pak, de Visser, & Shaw, 2018) have reported that people treat machine and human drivers equally. Kohn et al. (2018) found equivalent severity ratings of errors by machine and human drivers. Awad et al. (2020) observed that when only one driver (machine or human) makes an error, the erring machine and human are blamed equally, and that when both the machine and human make errors in cases of semi-AVs, the machine is attributed lower blame and responsibility (Note: the ratings of blame and responsibility attributed to the machine and the car company are aggregated in their study). Thus, Awad et al. (2020) concluded that people underreact to crashes caused by SDVs and that the human driver is blamed more than automation when the two share control of a semi-AV and both make errors.

These earlier findings, however, are challenged in simulator or vignette-based studies. In a simulator study, Waytz, Heafner, and Epley (2014) reported that an AV is blamed more than a human-driven vehicle (HDV) when the two are separately involved in an accident caused by another vehicle. Similarly, SDVs are blamed more than HDVs when they cause an accident and other road users' death (Hong, 2020; Sturgis, 2018); an automated truck is blamed more than a human-driven truck in a hypothetical collision with a tourist bus (without any cues on fault) (Dougherty, Stowell, Richards, & Ellen, 2018). Liu, Du, and Xu (2019), in observing people's biased response to hypothetical crashes caused by an SDV, identified a tendency to perceive a crash involving the SDV to be more severe than that involving an HDV, regardless of crash type (injury or fatality) or whether these crashes are caused by the involved SDV/HDV or not. Young and Monroe (2019) considered blame attribution and moral judgment when an SDV and a human driver are faced with a hypothetical moral dilemma, similar to Awad et al. (2020), but found that for identical decisions and outcomes, the SDV is judged as more blameworthy, less moral, and less trustworthy compared with the human driver.

Regarding semi-AVs, only Awad et al. (2020) conducted a study and found that in this case, the human is blamed more and assigned more responsibility than the automation counterpart when they are assumed to cause crashes. However, other works on other AVs (Hong, 2020; Liu et al., 2019; Waytz et al., 2014; Young & Monroe, 2019) indicate that lay people may overreact to crashes caused by automation; they tend to assign more blame and responsibility to the automation (including its creators) than to the human driver. Thus, we assumed that automation and the crash it causes are more negatively judged. We considered five judgment responses, namely, severity and acceptance (Liu et al., 2019), blame and responsibility (Waytz et al., 2014), and compensation (Siegrist & Sütterlin, 2014). In contrast to Awad et al. (2020), which examined crashes jointly caused by automation and human drivers, we focused on the crash caused by each. We formulated five hypotheses for testing:

H1−2:
: An automation-caused crash is perceived to be more severe (H1) and less acceptable (H2) compared with an equivalent human-caused one, regardless of crash severity.

H3−4:
: Automation in a semi-AV and its manufacturers are blamed more (H3) and assigned more responsibility (H4) compared with the human driver when they cause equivalent crashes.

H5:
Financial compensation to a victim in an automation-caused crash is judged to be higher than that in a human-caused crash.

2.3. Affect Heuristic

Affective responses to an object can directly influence people's evaluations, judgments, and decisions linked to the object, which is known as the "affect heuristic" in behavioral decision research (Finucane, Alhakami, Slovic, & Johnson, 2000; Slovic, Finucane, Peters, & MacGregor, 2007), "affect-as-information" model in social psychology (Schwarz & Clore, 1983), and "How-do-I-feel-about it?" heuristic in consumer research (Pham, 1998). Affective responses function in an affect-congruent direction; for instance, more positive affect will lead to more positive evaluation and judgment. Affect heuristic indicates the role of affect and emotions in risk perceptions, judgments, and risk-related behaviors (Loewenstein, Weber, Hsee, & Welch, 2001; Pachur, Hertwig, & Steinmann, 2012; Siegrist & Sütterlin, 2014; Slovic, 2000). For instance, negative affect triggered by an unexpected outcome determines its severity and acceptance rating (Siegrist & Sütterlin, 2014). People consult their affective response evoked by a cancer risk to decide how much money should be spent to avoid a single death from this cancer (Pachur et al., 2012). Liu et al. (2019) pointed out that owing to people's reliance on affect heuristic, they hold more negative affect toward SDVs (vs. HDVs) and toward the crashes involving SDVs (vs. those involving HDVs), leading to a higher perceived severity and lower acceptability of the crashes involving SDVs.

Legal judgments should be no exception. Although jurors and other legal decisionmakers are expected to be rational actors attempting to select the verdict that best applies the law to the facts of a case (Korobkin & Ulen, 2000), affect and emotions (particularly, fear and anger) can directly influence on judgments of legal responsibility and blame (Bornstein & Wiener, 2000; Bright & Goodman-Delahunty, 2006; Feigenson & Park, 2006; Slovic, 2000). For instance, gruesome photographs (vs. neutral or no photographs) elicit greater anger of mock jurors at the defendant, leading to an increased likelihood of conviction (Bright & Goodman-Delahunty, 2006). Similarly, lay people consult their affective responses to drive their decisions on blame and responsibility attributions.

Accordingly, differences between automation- and human-caused crashes in terms of severity and acceptance judgment, blame and responsibility attribution, and compensation judgment can be ascribed to people's different affective responses to these crashes. We further formulated five hypotheses:

H6−10:
Negative affect evoked by crashes involving semi-AVs mediates the relation between cause type (automation vs. human) and severity judgment of crashes (H6), acceptability of crashes (H7), blame attribution (H8), responsibility attribution (H9), and compensation judgment (H10).

2.4. Research Objectives

There are few direct comparisons of blame and responsibility attributed to automation versus human drivers in semi-AVs. We avoided looking at blame and responsibility judgments from legal, ethical, or philosophical perspectives, focusing instead on the perspective of lay people. As they will be final consumers of AVs, their opinions and preferences in these issues should be given serious consideration. Our first objective was to shed light on whether people respond differently to hypothetical crashes involving semi-AVs caused by human and automation in terms of severity judgment and acceptability of the crashes, blame and responsibility attribution to human and automation, and compensation to the victims of these crashes (see H1−H5). Given the existence of these differences, our second objective was to explore the psychological mechanism underlying them and to examine whether affective responses evoked by these crashes determine these differences (see H6−H10). Two studies were conducted to examine these hypotheses. Their data and results are available at the Open Science Framework (https://osf.io/g9kx5/?view_only=889699e5592b439cb55929ca2b685b72).

3. STUDY 1

In Study 1, we designed two vignette-based experiments with the manipulation of crash severity (injury in Experiment 1a and fatality in Experiment 1b). Study 1 investigated whether participants would perceive an automation-caused crash (vs. an equivalent human-caused crash) as more severe (H1) and less acceptable (H2), and if yes, whether

negative affect evoked by these crashes can account for differences in severity judgment (H6) and acceptability of the crashes (H7).

3.1. Methodology

A two (cause: human vs. automation) between-subjects design was adopted. In Experiment 1a, 240 students (120 women; 135 driving license holders; age: mean, $M = 21.9$ years, standard deviation, $SD = 1.7$) from a Chinese university were recruited and randomly assigned to one of the two groups (120 for each group).

In Experiment 1a, the participants first read the following introduction of semi-AVs: "In a semi-automated vehicle (also called conditionally automated vehicle), the automated driving system can perform all aspects of driving tasks, including controlling speed and steering. However, this system cannot handle all possible situations. The driver should continuously monitor the vehicle and roads. If the automated driving system cannot handle a situation, it will send a take-over request and the driver must retake control of the vehicle and perform manual driving." This introduction was adapted from Kyriakidis, Happee, and de Winter (2015). Next, they read a piece of text related to semi-AVs: "Assume a human driver driving a semi-automated vehicle with average driving safety performance of all human drivers. You are a passenger who is planning to take the driver's vehicle to your destination." Subsequently, they rated their levels of fear, anxiety, and trust in the semi-AV on three 10-point scales (1 = *very low*; 10 = *very high*) if they were required to ride in the semi-AV. Fear and anxiety measure the prior negative affect (Liu et al., 2019).

Next, the participants turned to a new page on which crash information was displayed: "On an urban road, this driver drove this semi-automated vehicle and took a passenger to a specific destination. However, owing to errors made by the automated driving system, a sudden traffic crash occurred and injured the passenger. The automated driving system did not send a take-over request," for the automation-caused crash, or "On an urban road, this driver drove this semi-automated vehicle and took a passenger to a specific destination. A sudden event occurred and the automated driving system sent a take-over request. However, owing to errors made by the human driver, the driver did not successfully take over the vehicle, which resulted in a traffic crash and injured the passenger," for the human-caused crash. After reading the crash information, the participants responded to the following questions: "What negative feeling did you experience because of the crash?" "How severe do you consider the crash to be?" and "How acceptable do you consider this kind of crashes to be?" (1 = *very low*; 10 = *very high*). These questions, adapted from previous studies (Liu et al., 2019; Siegrist & Sütterlin, 2014), measured the negative affect evoked by the crash, perceived severity of the crash, and acceptability of the crash, respectively. The participants also provided their ratings of fear, anxiety, and trust related to the semi-AV after reading the crash information. These responses were unrelated to the current study and thus were not reported here. The participants answered a question on responsibility attribution: "Who is the major party responsible for the crash?" (Three options were provided: the automated driving system, human driver, or neither of them). The crash vignette clearly indicated that the automated driving system or the human driver made mistakes. Thus, this question was intended to detect inattentive participants. The participants finally reported their sex, age, and whether they had a driving license, and then received a small present for their participation.

The measure and procedure design in Experiment 1b were exactly as in Experiment 1a, except that the crash resulted in the death of the passenger. Similarly, 240 students (120 women; 132 driving license holders; age: $M = 21.7$ years, $SD = 1.9$) were recruited and randomly assigned to one of the two groups (120 for each).

3.2. Results

Here we introduce the results of the two experiments separately. First, we examined the participants' response to the question of who is the major responsible party for the crash, which was initially used for attention check. The responses to this question were supposed to be the same between the automation- and human-caused crashes. In Experiment 1a, in the automation-caused crash, 108 participants regarded the automated system as the major responsible party, as expected, whereas four considered the human driver as the major responsible party, and three, neither; in the human-caused crash, 87 participants responded as expected, whereas 19 and 14 identified the automated system and neither as the major responsible party, respectively (see Table I). The correctness of the participants' responses was

Table I. Participants' Responses to the Question of "Who is the major party responsible for the crash?" in Experiments 1a and 1b

	Cause Type	The Major Party Responsible for the Crash		
		Automation	Human	None of Them
Experiment 1a (Injury)	Automation-caused	108	4	8
	Human-caused	19	87	14
Experiment 1b (Fatality)	Automation-caused	114	6	0
	Human-caused	13	104	3

dependent on injury cause ($\chi^2 = 12.1$, $p < 0.001$). In the chi-squared test, correct responses were coded as "1" and incorrect responses as "0." A similar trend was found in Experiment 1b (see Table I); the correctness of participants' responses was dependent on fatality cause ($\chi^2 = 5.0$, $p = 0.025$).

Thus, although this question was supposed to detect inattentive participants, it detected, to a large degree, an asymmetry in responsibility attribution: People are more likely to attribute the fault of the human in the human-caused crash to other agents than to attribute the fault of automation in the automation-caused crash to other agents.

3.2.1. Severity and acceptance judgment

Data of participants ($n = 195$) responding correctly to the question of responsibility attribution were analyzed. The participants in these two groups did not differ in terms of sex ($\chi^2 = 0.28$, $p = 0.598$), age ($F = 0.17$, $p = 0.680$), and possession of driving license ($\chi^2 = 0.16$, $p = 0.691$). Analysis of variance (ANOVA) tests were conducted. Before reading the injury crash information in Experiment 1a, the participants did not have different prior negative affect ($M_{\text{human}} = 5.61$, $M_{\text{auto}} = 5.50$, $F = 0.13$, $p = 0.722$, $\eta_p^2 = 0.001$) and trust in the semi-AV ($M_{\text{human}} = 5.40$, $M_{\text{auto}} = 5.43$, $F = 0.01$, $p = 0.935$, $\eta_p^2 < 0.001$). After reading the crash information, they reported higher negative affect evoked by the automation-caused crash ($M_{\text{human}} = 7.11$, $M_{\text{auto}} = 7.94$, $F = 9.40$, $p = 0.002$, $\eta_p^2 = 0.046$), perceived a higher severity of the crash ($M_{\text{human}} = 7.28$, $M_{\text{auto}} = 7.85$, $F = 5.00$, $p = 0.026$, $\eta_p^2 = 0.025$), and rated the acceptability of the crash lower ($M_{\text{human}} = 3.91$, $M_{\text{auto}} = 3.16$, $F = 11.71$, $p < 0.001$, $\eta_p^2 = 0.057$) (see Fig. 1). H1 and H2 were thus not rejected in Experiment 1a.

In Experiment 1b, the participants (218 responding correctly to the question of responsibility attribution) in these two groups did not differ in terms of sex ($\chi^2 = 0.02$, $p = 0.887$), age ($F = 1.74$, $p = 0.188$), and possession of driving license ($\chi^2 = 0.11$, $p = 0.743$). As in Experiment 1a, the participants reported nondifferent prior negative affect ($M_{\text{human}} = 5.29$, $M_{\text{auto}} = 5.52$, $F = 0.55$, $p = 0.458$, $\eta_p^2 = 0.003$) and trust in the semi-AV ($M_{\text{human}} = 5.21$, $M_{\text{auto}} = 5.06$, $F = 0.27$, $p = 0.603$, $\eta_p^2 = 0.001$), before reading the fatality crash information. After reading it, they reported higher negative affect evoked by the automation-caused crash ($M_{\text{human}} = 7.22$, $M_{\text{auto}} = 8.22$, $F = 18.48$, $p < 0.001$, $\eta_p^2 = 0.079$), perceived a higher severity of the crash ($M_{\text{human}} = 7.23$, $M_{\text{auto}} =$

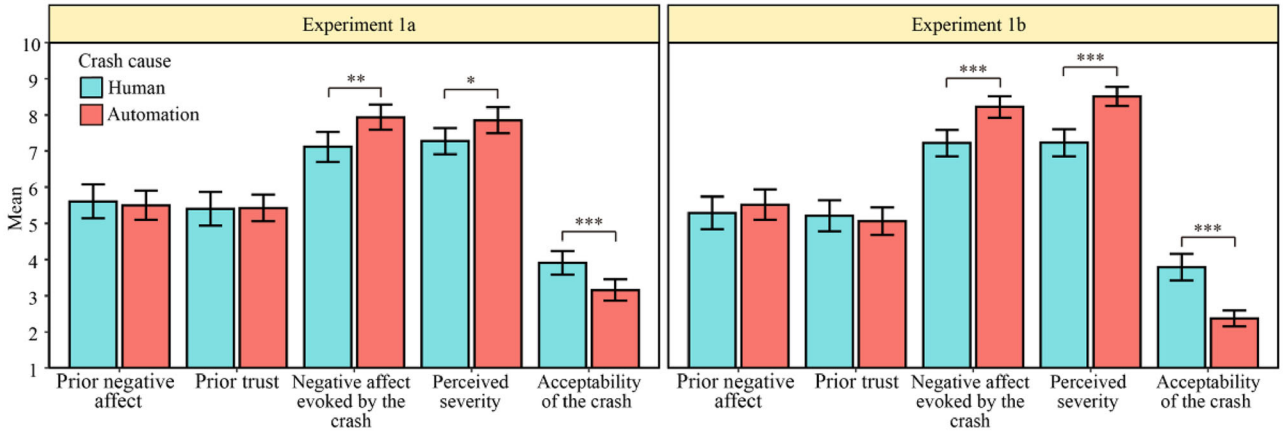

Fig 1. Means of five responses in Experiment 1a (left; injury crash) and Experiment 1b (right; fatality crash). $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Error bars = 2 standard error (*SE*).

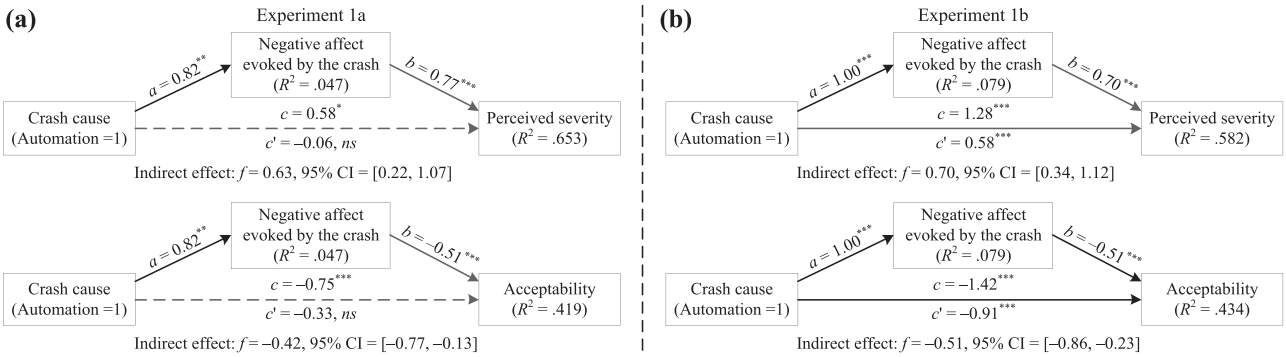

Fig 2. Results of mediation analysis in Experiments 1a (a; injury crash) and 1b (b; fatality crash). Nonstandardized coefficients are shown. Nonsignificant paths are shown as dotted lines. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; *ns*, nonsignificant.

8.51, $F = 32.03$, $p < 0.001$, $\eta_p^2 = 0.129$), and rated the acceptability of the crash lower ($M_{human} = 3.79$, $M_{auto} = 2.37$, $F = 44.88$, $p < 0.001$, $\eta_p^2 = 0.172$) (see Fig. 1). Thus, H1 and H2 were also confirmed in Experiment 1b.

3.2.2. Mediation Analysis

We adopted the PROCESS macro (Model 4) (Hayes, 2018) with 5,000 bootstrapped samples in mediation analysis. Cause type (human = 0, automation = 1) served as the independent variable, negative affect evoked by the crash, as the mediator, and perceived severity and acceptability of the crash, as separate dependent variables.

In Experiment 1a, the direct effect of cause type was nonsignificant after the inclusion of the mediator (on perceived severity: $c' = -0.06$, $p = 0.714$; on acceptability: $c' = -0.33$, $p = 0.061$) (see Fig. 2(a)). The bias-corrected 95% confidence intervals (CIs) of the indirect effect of cause type on both perceived severity and acceptability did not include zero. Thus, negative affect evoked by the injury crash fully mediated the relation between its cause type and the two dependent variables. Thus, Experiment 1a supported H6 and H7.

In Experiment 1b, the direct effect of cause type still kept as significant after the inclusion of the mediator (on perceived severity: $c' = 0.58$, $p < 0.001$; on acceptability: $c' = -0.91$, $p < 0.001$) (see Fig. 2(b)). The bias-corrected 95% CIs of the indirect effect of cause did not include zero. Thus, negative affect evoked by the fatality crash partially mediated the relation between its cause type and the two dependent variables. Thus, Experiment 1b also supported H6 and H7.

3.3. Summary

Experiments 1a and 1b yielded convergent findings. First, although our hypothetical scenarios clearly indicated which agent caused the crash, the number of participants attributing the cause to automation in the human-caused crash was higher than that doing so to the human driver in the automation-caused crash. Thus, automation may be more likely to be regarded as the major agent responsible for crashes caused by semi-AVs and blamed in these crashes, which might indicate an asymmetrical causality attribution in crashes related to semi-AVs. This bias seems to be in line with the "defensive attribution hypothesis" in social psychology (Shaver, 1970). Defensive attribution is the tendency of an observer to attribute the causes for a mishap to minimize their fear of being a victim or a cause in a similar situation. This self-protective motive influences responsibility and blame attributions (Burger, 1981). The observer will attribute more responsibility when this observer and the target of the blame share fewer similarities.

Second, the participants judged the automation-caused crash (vs. human-caused one) to be more severe and less acceptable, regardless of crash severity (injury or fatality), contrary to the findings by Awad et al. (2020) and Kohn et al. (2018) that people treat the errors of human and machine drivers equally. Our finding is consistent with previous research (Liu et al., 2019) in which a hypothetical crash involving a fully automated vehicle is more negatively judged than an equivalent crash involving a conventional vehicle.

Third, the negative affect triggered by the crash fully (Experiment 1a) or partially (Experiment 1b) mediated the relation between crash cause and negative evaluations of the crash. More specifically, compared with the human-caused crash, the automation-caused one triggered higher negative affect, which in turn led participants to produce more negative evaluations in an affect-congruent direction (i.e., higher severity rating and lower acceptability). This finding supports the affect heuristic (Finucane et al., 2000) and "affect-as-information" model (Schwarz & Clore, 1983).

4. STUDY 2

We conducted two vignette-based experiments (a young sample in Experiment 2a and a more diverse sample in Experiment 2b) to test whether participants ascribe more blame (H3) and responsibility (H4) to automation (and its creators) versus the human driver when each of them leads to an equivalent crash, whether they think the victim in the automation-caused crash should be compensated more (H5), and whether negative affect evoked by these crashes accounts for the differences in blame attribution (H8), responsibility attribution (H9), and compensation judgment (H10).

4.1. Methodology

A two (cause: human vs. automation) between-subjects design was adopted. The questionnaires used in Experiments 2a and 2b shared similarities. In Experiment 2a (an offline survey), 240 college students (121 women; 143 driving license holders; age: $M = 22.4$ years, $SD = 2.1$) from a Chinese university were recruited and randomly assigned to one of the two groups (120 for each group). The participants in these two groups did not differ in sex ($\chi^2 = 0.02$, $p = 0.897$), age ($F = 1.49$, $p = 0.223$), and possession of driving license ($\chi^2 = 0.43$, $p = 0.511$). The participants first read an introduction of semi-AVs (see Study 1). Next, they read the following crash information: "On a highway road, a driver drove this semi-automated vehicle and took a passenger to a specific destination. During the trip, a car ahead stopped suddenly, at the moment of which the driver controlled the semi-automated vehicle. The automated driving system immediately sent a collision warning. If the driver had been able to step on the brake and avoid the car ahead immediately, then the driver could have avoided an accident. However, this driver was inattentive at that time and did not brake and avoid the car ahead in a timely manner. Finally, a crash occurred and caused the death of the passenger," for the human-caused crash, or "On a highway road, a driver drove this semi-automated vehicle and took a passenger to a specific destination. During the trip, a car ahead stopped suddenly, at the moment of which the automated driving system controlled the semi-automated vehicle. If the automated driving system had been able to engage the brake and avoid the car ahead immediately, it could have avoided an accident. However, the automated driving system malfunctioned at that time and did not engage the brake and avoid the car ahead nor send a take-over request in a timely manner. Finally, a crash occurred and caused the death of the passenger," for the automation-caused crash.

After this, the participants answered three questions. Regarding the human-caused crash, the question for blame attribution was "To what extent do you think the driver should be blamed for this crash?" (1 = *very little*, 10 = *very much*). The question for responsibility attribution was "To what extent do you think the driver caused the death of this passenger in this crash?" (1 = *very little*, 10 = *very much*). The question for compensation judgment was "Assume the casualty was a middle-aged man, an urban resident. How much do you think should his family receive as financial compensation for his death?" adapted from Siegrist and Sütterlin (2014). According to a statistic in Tianjin City in China, the average financial compensation for the death of an urban resident in traffic accidents is about CNY 800,000 (https://www.peichang.cn/detail/id21818.html). We provided four options for the compensation question: CNY 600,000, CNY 800,000, CNY 1 million, and CNY 1.2 million. In the automation-caused crash, the responsible party was replaced by "the automated driving system and its manufacturer" in the questions for blame and responsibility attributions, similar to Awad et al. (2020). Finally, the participants reported their sex, age, and whether they held a driving license.

In Experiment 2b, more diverse participants ($n = 325$) were recruited online through social media tools (119 women; 205 driving license holders; age: $M = 28.9$ years, $SD = 6.5$, min = 20, max = 58). They were instructed to choose the human- ($n = 169$) or automation-caused crash scenario ($n = 156$) according to the parity of the last digit of their cell phone number. The participants in these two groups did not differ in sex ($\chi^2 = 0.07$, $p = 0.796$), age ($F = 0.05$, $p = 0.829$), and possession of driving license ($\chi^2 = 0.10$, $p = 0.747$). After reading the crash information, they first responded to the question "What negative feeling did you experience because of the crash?" (1 = *very low*; 10 = *very high*) (see Study 1). Next, they responded to the three questions related to blame and responsibility attributions and compensation judgment as in Experiment 2a.

4.2. Results

4.2.1. Blame and Responsibility Attribution

In Experiment 2a, ANOVA indicated that automation and its manufacturer in the automation-caused crash were blamed more ($M_{human} = 7.47$, $M_{auto} = 8.03$, $F = 7.04$, $p = 0.009$, $\eta_p^2 = 0.029$) and assigned more responsibility ($M_{human} = 7.53$, $M_{auto} = 8.11$, $F = 8.44$, $p = 0.004$, $\eta_p^2 = 0.034$) than the human driver in the human-caused crash (see Fig. 3), supporting H3 and H4.

In Experiment 2b, similarly, automation and its manufacturer in the automation-caused crash were blamed more ($M_{human} = 7.42$, $M_{auto} = 7.92$, $F = 5.96$, $p = 0.015$, $\eta_p^2 = 0.018$) and assigned more responsibility ($M_{human} = 7.22$, $M_{auto} = 7.90$, $F = 10.37$, $p = 0.001$, $\eta_p^2 = 0.031$), supporting H3 and H4. In addition, the automation-caused crash evoked more negative affect ($M_{human} = 7.04$, $M_{auto} = 7.81$, $F = 14.49$, $p < 0.001$, $\eta_p^2 = 0.043$).

4.2.2. Compensation Judgment

Results of the chi-squared tests indicated that student participants in Experiment 2a did not show different compensation judgments for the two crashes ($\chi^2 = 3.30$, $p = 0.348$); however, in Experiment 2b, which had a more diverse sample, the victim in the automation-caused crash was judged as deserving to be compensated more compared with that in the human-caused crash ($\chi^2 = 10.22$, $p = 0.017$) (see Fig. 4). Thus, H5 was supported in Experiment 2b but not in Experiment 2a.

4.2.3. Mediation Analysis

Mediation analysis revealed that the direct effect of cause was not significant after the inclusion of negative affect (on blame attribution: $c' = 0.03$, $p = 0.865$; on responsibility attribution: $c' = 0.25$, $p = 0.167$) (see Figs. 5(a) and (b)). Negative affect fully mediated the relation between crash cause and attributions of blame and responsibility in Experiment 2b, supporting H8 and H9.

Mediation analysis involving categorical dependent variables is complex. Research has provided ways to run mediation analysis when the dependent variable is binary (MacKinnon, 2008). We coded compensation judgment as a binary variable with two values: 0 (CNY 600,000 and 800,000) and 1 (CNY 1 million and 1.2 million). We followed the procedure proposed by MacKinnon (2008), which is based on the Sobel z-test (Sobel, 1982), and ran logistic regression with compensation judgment as the binary dependent variable and cause type and negative affect as its predictors. The mediated effect (i.e., indirect effect) was calculated as the product of a from ordinary

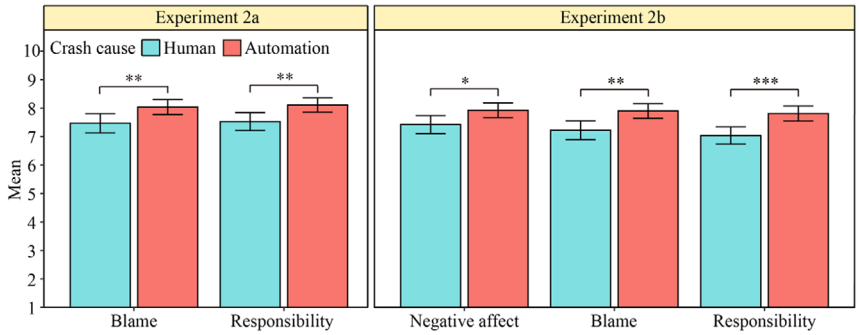

Fig 3. Means of responses in Experiment 2a (left; a student sample) and Experiment 2b (right; a more diverse sample). $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. Error bars = 2 *SE*.

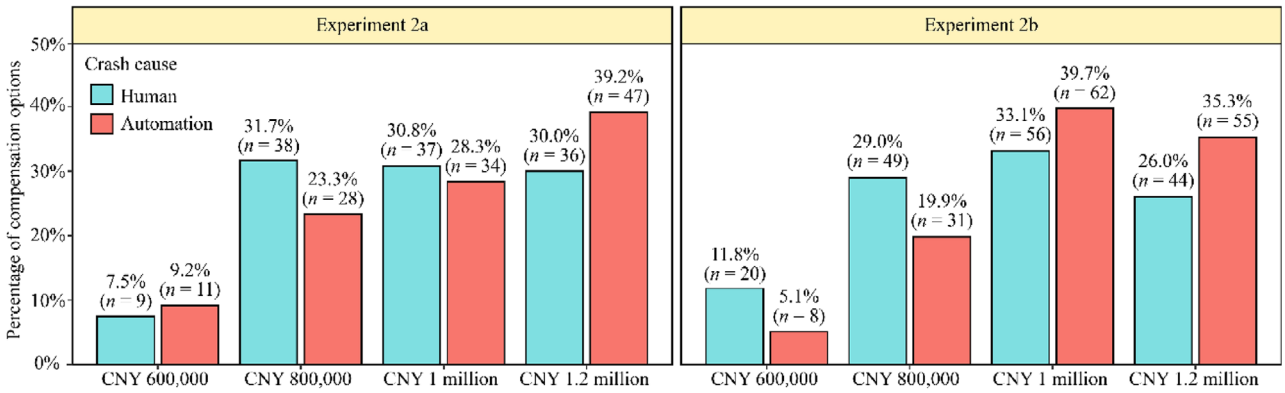

Fig 4. Compensation judgment (Chinese Yuan, CNY) in the human- and automation-caused crashes in Experiment 2a (left; a student sample) and Experiment 2b (right; a more diverse sample).

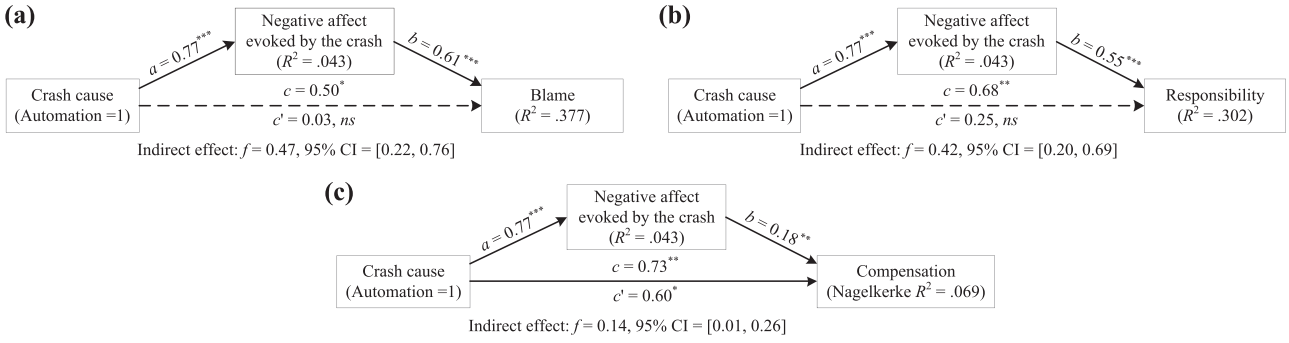

Fig 5. Results of mediation analysis in Experiment 2b. Nonstandardized coefficients are shown. Nonsignificant paths are shown as dotted lines. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; *ns*, nonsignificant.

least squares regression and *b* from logistic regression in Fig. 5(c), which was 0.14 (95% CI [0.01, 0.26]). The direct effect of cause on compensation judgment remained significant ($c' = 0.60$, $p = 0.015$) after the inclusion of negative affect, implying that negative affect partially mediated the influence of cause type on compensation judgment (see Fig. 5(c)). Thus, H10 was supported.

4.3. Summary

Both vignette-based Experiments 2a and 2b indicated that participants attributed more blame and responsibility to automation (and its creators) than to the human driver when the crashes caused by them resulted in equivalent negative outcomes in human–automation cooperation. In other comparative studies related to SDVs and human drivers, participants

also attribute more blame and responsibility to SDVs than to human drivers when both cause identical accidents (Dougherty et al., 2018; Hong, 2020; Young & Monroe, 2019). An unfair compensation judgment was observed in our study: the participants judged that the victim in the automation-caused crash should be compensated more financially than that in the human-caused one in Experiment 2b. Although the student participants in Experiment 2a did not show this pattern, the results from more diverse participants in Experiment 2b seemed more convincing as these participants could be expected to know more on financial compensation in actual traffic crashes. Following affect heuristic (Finucane et al., 2000), we explained that blame attribution, responsibility attribution, and compensation judgment in cases of automation causing a crash were amplified via higher activation of negative affect in this crash.

5. GENERAL DISCUSSION

5.1. Blame Attribution Asymmetry

Studies 1 and 2 measured different responses but reported convergent findings. In Study 1, a hypothetical automation-caused crash was judged to be more severe and less acceptable than an equivalent hypothetical human-caused crash, regardless of crash severity (injury or fatality). In Study 2, the participants ascribed more blame and responsibility to automation and its creators than the human driver who caused an equivalent crash. These findings were expected, to a certain degree, as there is indirect evidence supporting them (Dougherty et al., 2018; Hong, 2020; Sturgis, 2018; Waytz et al., 2014; Young & Monroe, 2019). Two other findings were entirely novel. First, we asked the participants to judge the major responsible party for the crash, the purpose of which was for checking their attention, and found that automation was more likely to be judged as the major responsible party for crashes in Study 1. Second, the results of Experiment 2b in Study 2 indicated that a victim in an automation-caused crash should be awarded more compensation than that in a human-caused one.

Thus, we provide previously unidentified evidence of a bias, which we called the "blame attribution asymmetry" and defined as a tendency for people to blame automation more and judge automation-caused crashes more negatively. This bias is not in line with certain vignette-based studies claiming that people treat the errors of human and machine drivers similarly (Kohn et al., 2018) and that less blame is attributed to the machine versus the human driver when both make errors in semi-AVs during their cooperation (Awad et al., 2020). Rather, it is aligned with findings for other types of AVs (e.g., fully automated vehicles) (Dougherty et al., 2018; Hong, 2020; Liu et al., 2019; Waytz et al., 2014; Young & Monroe, 2019). To the best of our knowledge, we provide the first evidence demonstrating people's overreaction to a crash caused by a machine driver and to the machine driver in semi-AVs.

The blame attribution asymmetry is mirrored by research on the diverging safety requirements for AVs and human drivers. People will have higher safety requirements for AVs when they have to entrust their lives to AVs (Shladover & Nowakowski, 2019; Waycaster, Matsumura, Bilotkach, Haftka, & Kim, 2018). Indeed, empirical studies have indicated that the acceptable safety of SDVs would be four to five times that of human drivers (Liu, Wang, & Vincent, 2020; Liu, Yang, & Xu, 2019) and that people require SDVs to be safer than their own perceived ability to drive safely (Nees, 2019).

When automation and human driver in semi-AVs were assumed to cause identical accidents, automation and its creators were attributed more responsibility in our study and less responsibility in Awad et al.'s study (2020) compared with the human driver. Awad et al. (2020) designed accident scenarios as moral dilemmas in which the automation or human driver has to distribute the unavoidable harm. However, we designed normal accident scenarios. Young and Monroe (2019) similarly considered moral dilemmas but reported findings contrary to those in Awad et al. (2020). Young and Monroe (2019) found that when SDVs and human drivers make identical moral decisions and cause identical accidents, SDVs are judged as more blameworthy and less moral compared with human drivers. Therefore, the difference in accident scenarios might not account for the different findings in Awad et al.'s and our studies. There must be other reasons for them, which warrant attention.

5.2. Affect Heuristic for Explaining the Blame Attribution Asymmetry

Affect, emotions, and feelings can directly guide people's evaluation, judgments, decision, and behaviors (Lerner, Li, Valdesolo, & Kassam, 2015; Loewenstein et al., 2001; Pachur et al., 2012; Pham,

1998; Schwarz & Clore, 1983). We extended the affect heuristic model (Finucane et al., 2000; Slovic et al., 2007), which suggests that people might rely on affect while arriving at judgments of risks and benefits, to explain the blame attribution asymmetry. More specifically, this bias would be a result of higher negative affect triggered by the automation-caused crash versus the human-caused one. Other affect-induced biases have also been identified in psychological and behavioral research (Gigerenzer, 2004; Lerner et al., 2015). Research has used the concept of mind perception (Bigman & Gray, 2018; Waytz et al., 2014) or the social and institutional roles of human and autonomous agents (Malle et al., 2019) to account for people's asymmetric responses to them. The role of affect and emotions is largely ignored. Our theoretical contribution is to confirm the affect heuristic as a theoretical account.

Negative affect partially or fully mediated the relations of crash cause with severity and acceptability judgments, blame and responsibility attributions, and compensation judgment. These effects are indirectly supported by previous studies. Siegrist and Sütterlin (2014) reported that negative affect associated with a hazard leads it to be more negatively evaluated in terms of severity judgment and public acceptability. Liu et al. (2019) found that higher prior negative affect tagged with an SDV (vs. a human-driven vehicle) intensifies people's negative affect evoked by a crash involving the SDV (vs. a crash involving the human-driven vehicle), leading to higher perceived severity and lower acceptability of the crash. Higher negative affect and emotions (e.g., anger) can amplify attributions of legal responsibility and blame (Bright & Goodman-Delahunty, 2006; Feigenson, 2010; Feigenson & Park, 2006). Pachur et al. (2012) reported that people's negative affect from a specific risk determines how much money they think the government should spend to avoid a fatality caused by the risk, indirectly supporting the role of negative affect on the relation between crash cause and compensation payment in our study.

5.3. Policy Implications

Understanding public responses to human and machine drivers may determine how the public is treated during the deployment of semi-AVs. Studies have argued that people have nondifferent responses to negative outcomes attributable to human and automation (Kohn et al., 2018) and that more blame is attributed to the human if the joint decision of human and automation fails (Awad et al., 2020). Thus, Awad et al. (2020) underscored the public's underreaction to automation's malfunctions. Contrary to Awad et al.'s argument, our study, and also previous work (Dougherty et al., 2018; Hong, 2020; Liu et al., 2019; Young & Monroe, 2019), clearly indicated people's overreaction to automation's malfunctions and its negative outcomes. Automation-caused crashes can be expected to produce more psychological costs. Considering that perceived severity is linked with protective intentions and behavior (Sheeran, Harris, & Epton, 2014), automation-caused crashes will deter more people from adopting semi-AVs.

As people are likely to overreact to automation-caused crashes, society should be cautious with the introduction policies of semi-AVs. Policy researchers (Kalra & Groves, 2017) are promoting a less stringent policy of allowing AVs on public roads as long as they are safer than the average human driver. Although such a policy could be beneficial to society in the long run, it might backfire, as under this policy, people would find that crashes involving AVs are as many as those involving conventional vehicles, the higher psychological costs of which might prevent people from adopting AVs. Thus, policymakers and regulators should be aware of people's overreaction to crashes involving AVs when they set policies for deploying and regulating AVs.

Our findings offer insight on attributions of blame and responsibility. Arguably, semi-AVs are not fully AVs and their autonomy is limited; meanwhile, a human driver in a semi-AV is supervising the driving and would take over control of the vehicle when requested, and thus, the human should be regarded as the responsible party for unexpected outcomes during the drive (Nyholm, 2018). However, subjective autonomy is more important than objective autonomy, and autonomy and moral responsibility are more matters of perception (Bigman et al., 2019). Our participants tended to attribute responsibility to automation and blame it and its creators more when it caused a crash.

Our findings on compensation decision might be of particular value to policymakers interested in issues related to financial compensation of victims injured or killed by machine drivers. Stakeholders are keenly interested in insurance and liability in the transformative age of AVs (Anderson, Kalra, Stanley, & Morikawa, 2018). According to our findings, they might need to consider the possibility that to

lay people, victims of AV crashes should be compensated more than commonly calculated. For instance, Germany has enacted a bill legalizing AVs with major modifications, one of which is the maximum liability limits under the Road Traffic Act, which has been doubled: the maximum EUR 5 million is increased to EUR 10 million for death or injury and the maximum EUR 1 million is increased to EUR 2 million for damage to property (Burianski & Theissen, 2017).

Our findings also imply that this affect-induced bias could prevent the public from obtaining the benefits from well-designed autonomous agents. The bias associated with semi-AVs must be reduced and mitigated, given that crashes are unavoidable. Certain strategies, such as direct experience (Liu, Xu, & Zhao, 2019), have been suggested to reduce the negative affect and nurture a positive image linked to AVs. In addition, informing the public about their affect-induced bias to automation and the risk of being averse to automation could be an effective and inexpensive intervention, as suggested by Gigerenzer (2004).

5.4. Limitations and Future Directions

Several limitations are noted, and corresponding future directions are suggested. First, the usage of unrepresentative samples from only one country could constrain the generalizability of our results. Replications with representative samples from multiple countries would confirm the universality of this blame attribution asymmetry. Second, we recruited participants without actual experience in driving semi-AVs, and their responses to automation and automation-caused crashes could have been influenced by an inaccurate perception of semi-AVs. Further research can consider inviting participants with experience in driving semi-AVs. Third, to furnish a parsimonious focus on testing the hypotheses, we did not consider the influence of specific affect and emotions (e.g., anger). Further works can investigate the specific affect and emotions that contribute the most to the affect-induced bias. Fourth, we designed each agent (automation and human) to have sole responsibility in hypothetical accident scenarios and ignored their responsibility in crashes attributable to their joint faulty decisions and actions. Other scenarios involving joint decisions or actions and blurring the responsibility between the human and automation driver deserve particular attention. Fifth, we explained the blame attribution asymmetry only through the lens of the affect heuristic. The role of other factors (e.g., mind perception and perceived dissimilarity between human and machine drivers) on blame attribution in semi-AVs warrants attention. Sixth, although vignettes with hypothetical scenarios are widely used in AV studies involving unwanted outcomes (e.g., our current study), a vignette-based design might lack ecological validity. As such, the generalizability of results from vignette-based studies should be considered with caution. Future studies can use a more vivid research design (e.g., video animation and virtual reality) to demonstrate to participants how a semi-AV works and how a crash is caused by the human or automation driver, and then to survey participants' responses to the crash and the agent responsible for the crash. Seventh, the participants' responses (e.g., negative affect) to our vignettes and questions could be biased owing to social desirability (or demand effects). Future studies can consider objective techniques to reduce this bias (e.g., measuring negative affect through facial emotion recognition). Finally, future studies can consider an open-ended question to measure the financial compensation that is deemed fair for victims in crashes.

6. CONCLUSIONS

As autonomous agents permeate the various sectors of society, their cooperation with human operators will be fundamental for improving people's quality of life and work. Although favorable outcomes from cooperation are highly desired, negative outcomes are unavoidable. People's responses to the negative outcomes of human–automation cooperation must be elucidated. The present study concentrated on the application of semi-AVs, which may operate on public roads very soon. In our effort to theorize people's asymmetric responses to crashes caused by each driver and the underlying psychological mechanism, we identified a bias, namely, blame attribution asymmetry, which refers to the tendency among people to ascribe more blame and responsibility to automation than to the human driver when each caused equivalent crashes and to judge the automation-caused crashes more harshly and negatively. This bias indicates people's overreaction to automation-caused crashes, probably owing to the higher negative affect evoked by these crashes. This work can be extended to scenarios involving the ubiquitous cooperation of other autonomous agents (e.g., AI, robots) and humans to examine the

generalizability of this affect-induced blame attribution asymmetry.

ACKNOWLEDGMENT

This research was funded by the National Natural Science Foundation of China (Project no. 72071143).

REFERENCES

Anderson, J. M., Kalra, N., Stanley, K. D., & Morikawa, J. (2018). *Rethinking insurance and liability in the transformative age of autonomous vehicles*. Santa Monica, CA: RAND Corporation..

Associated Press. (2018). Arizona Governor suspends uber from autonomous testing. *U.S. News*. Retrieved from https://www.usnews.com/news/technology/articles/2018-03-26/arizona-governor-suspends-uber-from-autonomous-testing

Atiyeh, C. (2015). Volvo will take responsibility if its self-driving cars crash. *Car and Driver*. Retrieved from https://www.caranddriver.com/news/a15352720/volvo-will-take-responsibility-if-its-self-driving-cars-crash/

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., … Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*, 134–143.

Banerjee, S. S., Jha, S., Cyriac, J., Kalbarczyk, Z. T., & Iyer, R. K. (2018). Hands off the wheel in autonomous vehicles? A systems perspective on over a million miles of field data. Paper presented at 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). Luxembourg City, Luxembourg.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, *23*(5), 365–368.

Bornstein, B. H., & Wiener, R. L. (2000). Emotion and the law: A field whose time has come. In B. H. Bornstein & R. L. Wiener (Eds.), *Emotion and the law: Psychological perspectives* (pp. 1–12). New York, NY: Springer.

Bright, D. A., & Goodman-Delahunty, J. (2006). Gruesome evidence and emotion: Anger, blame, and jury decision-making. *Law and Human Behavior*, *30*(2), 183–202.

Burger, J. M. (1981). Motivational biases in the attribution of responsibility for an accident: A meta-analysis of the defensive-attribution hypothesis. *Psychological Bulletin*, *90*(3), 496–512.

Burianski, M., & Theissen, C. M. (2017). Germany permits automated vehicles. Retrieved from https://www.whitecase.com/publications/article/germany-permits-automated-vehicles

Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence*, *30*(8), 748–757.

Conger, K. (2020). Driver charged in Uber's fatal 2018 autonomous car crash. Retrieved from https://www.nytimes.com/2020/09/15/technology/uber-autonomous-crash-driver-charged.html

Dougherty, S., Stowell, J., Richards, A., & Ellen, P. (2018). Will automated trucks trigger the blame game and socially amplify risks? Paper presented at Engaged Management Scholarship Conference. Philadelphia, PA.

Favarò, F. M., Nader, N., Eurich, S. O., Tripp, M., & Varadaraju, N. (2017). Examining accident reports involving autonomous vehicles in California. *Plos One*, *12*(9), e0184952. https://doi.org/10.1371/journal.pone.0184952

Feigenson, N. (2010). Emotional influences on judgments of legal blame: How they happen, whether they should, and what to do about it. In B. H. Bornstein & R. L. Wiener (Eds.), *Emotion and the law: Psychological perspectives* (pp. 45–96). New York: Springer.

Feigenson, N., & Park, J. (2006). Emotions and attributions of legal responsibility and blame: A research review. *Law Human Behavior*, *30*(2), 143–161.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, *13*(1), 1–17.

Gigerenzer, G. (2004). Dread risk, September 11, and fatal traffic accidents. *Psychological Science*, *15*(4), 286–287.

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97–103.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed). London, UK: The Guilford Press.

Hevelke, A., & Nida-Rümelin, J. (2015). Responsibility for crashes of autonomous vehicles: An ethical analysis. *Science and Engineering Ethics*, *21*(3), 619–630.

Hong, J. W. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction*, *36*(18), 1768–1774.

Kahn, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., … Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction. Boston, MA.

Kalra, N., & Groves, D. G. (2017). *The enemy of good: Estimating the cost of waiting for nearly perfect automated vehicles*. Santa Monica, CA: RAND Corporation.

Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. Proceedings of 15th IEEE International Symposium on Robot and Human Interactive Communication. Hatfield, UK.

Kohn, S. C., Quinn, D., Pak, R., de Visser, E. J., & Shaw, T. H. (2018). Trust repair strategies with self-driving vehicles: An exploratory study. Proceedings of the Human Factors and Ergonomics Society 2018 Annual Meeting. Philadelphia, PA.

Komatsu, T. (2016). Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds. Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). New Zealand.

Korobkin, R. B., & Ulen, T. S. (2000). Law and behavioral science: Removing the rationality assumption from law and economics. *California Law Review*, *88*(4), 1501–1144.

Kyriakidis, M., Happee, R., & de Winter, J. C. F. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F: Traffic Psychology and Behaviour*, *32*, 127–140.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*(1), 799–823.

Liu, P., Du, Y., & Xu, Z. (2019). Machines versus humans: People's biased responses to traffic accidents involving self-driving vehicles. *Accident Analysis & Prevention*, *125*, 232–240.

Liu, P., Wang, L., & Vincent, C. (2020). Self-driving vehicles against human drivers: Equal safety is far from enough. *Journal of Experimental Psychology: Applied*, *26*(4), 692–704.

Liu, P., Xu, Z., & Zhao, X. (2019). Road tests of self-driving vehicles: Affective and cognitive pathways in acceptance formation. *Transportation Research Part A: Policy and Practice*, *124*, 354–369.

Liu, P., Yang, R., & Xu, Z. (2019). How safe is safe enough for self-driving vehicles? *Risk Analysis*, *39*(2), 315–325.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin*, *127*(2), 267–286.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Lawrence Erlbaum Associates.

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

Malle, B. F., Thapa Magar, S., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In I. A. Ferreira, J. S. Sequeira, G. S. Virk, E. E. Kadar, & O. Tokhi (Eds.), *Robots and well-being* (pp. 111–133). Cham, Switzerland: Springer.

Maric, P. (2017). Audi to take full responsibility in event of autonomous vehicle crash. *Car Advice*. Retrieved from http://www.caradvice.com.au/582380/audi-to-take-full-responsibility-in-event-of-autonomous-vehicle-crash/

Motamedi, S., Wang, P., Zhang, T., & Chan, C.-Y. (2020). Acceptance of full driving automation: Personally owned and shared-use concepts. *Human Factors*, *62*(2), 288–309.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103.

National Highway Traffic Safety Administration. (2016). *Federal automated vehicles policy: Accelerating the next revolution in roadway safety*. Washington, DC: National Highway Traffic Safety Administration, U.S. Department of Transportation

National Transportation Safety Board. (2019). *Collision between vehicle controlled by developmental automated driving system and pedestrian Tempe, Arizona March 18, 2018*. Highway Accident Report. Report no. NTSB/HAR-19/03. Washington, DC: National Transportation Safety Board

Nees, M. A. (2019). Safer than the average human driver (who is less safe than me)? Examining a popular safety benchmark for self-driving cars. *Journal of Safety Research*, *69*, 61–68.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*(4), 1201–1219.

Pachur, T., Hertwig, R., & Steinmann, F. (2012). How do people judge risks: Availability heuristic, affect heuristic, or both? *Journal of Experimental Psychology: Applied*, *18*(3), 314–330.

Pham, M. T. (1998). Representativeness, relevance, and the use of feelings in decision making. *Journal of Consumer Research*, *25*(2), 144–159.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*(3), 513–523.

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 1–16.

Shaver, K. G. (1970). Defensive attribution: Effects of severity and relevance on the responsibility assigned for an accident. *Journal of Personality and Social Psychology*, *14*(2), 101–113.

Sheeran, P., Harris, P. R., & Epton, T. (2014). Does heightening risk appraisals change people's intentions and behavior? A meta-analysis of experimental studies. *Psychological Bulletin*, *140*(2), 511–543.

Shladover, S. E., & Nowakowski, C. (2019). Regulatory challenges for road vehicle automation: Lessons from the California experience. *Transportation Research Part A: Policy and Practice*, *122*, 125–133.

Siegrist, M., & Sütterlin, B. (2014). Human and nature-caused hazards: The affect heuristic causes biased decisions. *Risk Analysis*, *34*(8), 1482–1494.

Slovic, P. (2000). Rational actors and rational Fools: The influence of affect on judgment and decision-making. *Roger Williams University Law Review*, *16*(1), 163–211).

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, *177*(3), 1333–1352.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312.

Society of Automotive Engineers. (2018). *Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems* (pp. 3016–2018). Washington, DC: SAE International

Sturgis, J. (2018). *The effect of anthropomorphism on blame attribution in autonomous cars* (Master Thesis, Edinburgh Napier University, Edinburgh, Scotland).

Tesla. (2016). *A tragic loss*. Tesla. Retrieved from https://www.tesla.com/blog/tragic-loss

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. robot agents. Proceedings of 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). New York.

Waycaster, G. C., Matsumura, T., Bilotkach, V., Haftka, R. T., & Kim, N. H. (2018). Review of regulatory emphasis on transportation safety in the United States, 2002–2009: Public versus private modes. *Risk Analysis*, *38*(5), 1085–1101.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology*, *85*, 103870.