Are autonomous vehicles blamed differently?

Darko Stojilović[1] (darko.stojilovic.22@ucl.ac.uk), **Matija Franklin**[1,7] (matija.franklin@ucl.ac.uk),
Bertram Malle[2] (bfmalle@brown.edu), **Carlos Fernandez-Basso**[1,3] (cjferba@decsai.ugr.es),
Edmond Awad[4,5,6] (e.awad@exeter.ac.uk), **David Lagnado**[1] (d.lagnado@ucl.ac.uk)

[1] Department of Experimental Psychology, University College London, London, UK
[2] Department of Cognitive, Linguistic & Psychological Sciences, Brown University, Providence, RI, USA
[3] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[4] The Oxford Uehiro Centre for Practical Ethics, University of Oxford, Oxford, UK
[5] Department of Economics and Institute for Data Science and AI, University of Exeter, Exeter, UK
[6] Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany
[7] AI Objectives Institute, San Francisco, CA, USA

Abstract

This study investigates how people assign blame to autonomous vehicles (AVs) when involved in an accident. Our experiment ($N = 2647$) revealed that people placed more blame on AVs than on human drivers when accident details were unspecified. To examine whether people assess major classes of blame-relevant information differently for AVs and humans, we developed a causal model and introduced a novel concept of prevention effort, which emerged as a crucial factor for blame judgement alongside intentionality. Finally, we addressed the "many hands" problem by exploring how people assign blame to entities associated with AVs and human drivers, such as the car company or an accident victim. Our findings showed that people assigned high blame to these entities in scenarios involving AVs, but not with human drivers. This necessitates adapting a model of blame for AVs to include other agents and thus allow for blame allocation "outside" of autonomous vehicles.

Keywords: blame attribution; autonomous vehicles; causal models

Introduction

In May 2023, an artificial intelligence (AI) generated a fake image of an explosion near the Pentagon and caused a market flash crash. This capacity of AI to cause harm is not a recent revelation. Autonomous military drones that can select targets on their own and engage in lethal action without human control have been deployed in warfare for some time now (Konert & Balcerzak, 2021). Autonomous vehicles (AVs), too, have been involved in road accidents, leading to property damage, bodily injury, and even fatalities (Favarò et al., 2017). In a 2023 Alan Turing Institute poll, nearly three quarters of British citizens voiced their concerns over AI being used in driverless vehicles and autonomous weaponry[1].

To effectively establish formal-legal guidelines for assigning responsibility and liability in cases where AI causes harm, it is crucial to understand how people informally assign blame in such scenarios. Blame judgements rely on social-cognitive processes, such as intentionality judgements, mental states inferences, and causal reasoning (Alicke, 2000; Cushman, 2008; Lagnado & Channon, 2008; Malle, Guglielmo, & Monroe, 2014). Although at least some AI systems make autonomous decisions and bring about the same causal effects as humans do, ascribing mental states to AI is challenging (Scheutz & Malle, 2021; Wallach, Franklin, & Allen, 2010). An even deeper problem is whether people think of artificial intelligence systems as moral agents at all and as proper targets of blame (Banks, 2019).

In complex real-world situations, assigning blame is particularly challenging as multiple agents may contribute to the outcome – the problem of Many Hands (Van de Poel & Nihlén Fahlquist, 2012). Autonomous agents further complicate this, because their contributions are intertwined with those of its developers, users, and the potential victims of its behaviour (Nyholm, 2018). This intricacy can cause responsibility gaps, where the uncertainty over who is to blame can leave nobody taking the blame (Danaher, 2016; De Jong, 2020). In fact, human agents may escape accountability by insisting that they delegated responsibility onto the autonomous system (Danaher, 2019; Parlangeli et al., 2023).

This paper concentrates on autonomous vehicles. We compare the blameworthiness of AVs to that of a human driver in the same context, and explore the role of other agents involved. The central focus revolves around the concept of blame, which we operationalise as a continuous judgement and examine through a causal model.

Theoretical Framework

The path model of blame by Malle et al. (2014) posits several classes of information that determine how much blame an agent deserves, and we use this model as a starting point to develop our research question. Are these classes of information evaluated differently depending on whether the agent is an autonomous vehicle or a human? Intentionality is pivotal. Actions deemed intentional are typically assigned more blame than those considered unintentional (Lagnado & Channon, 2008), and intentionality judgements guide additional information processing—whether to look more for mental states or for causal-counterfactual information (Cushman, 2008; Monroe & Malle, 2017).

[1] turing.ac.uk/news/publications/how-do-people-feel-about-ai

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Desires constitute one important mental state that people consider when blaming intentional actions (Cushman, 2008). If an agent's desire specifically motivates a harmful action, blame will be considerable. Still, the presence of other desires can mitigate blame, such as wanting to save someone's life by yanking them back onto the sidewalk from an approaching speeding car. Desires of autonomous agents can be thought of as goals – the outcomes they aim to achieve in the environment (Ashton, 2022a; 2022b). But inferring artificial agents' specific goals can be difficult (Castelvecchi, 2016).

In situations where an agent unintentionally caused harm, people process other classes of information. They consider whether (a) the agent had an obligation to prevent the outcome, (b) the agent foresaw it, and (c) could have prevented it (Malle et al., 2014). Agents are assigned more blame for outcomes they should have prevented, foresaw, and could have prevented (Gerstenberg et al., 2011; Lagnado & Channon, 2008; Monroe & Malle, 2019; Weiner, 1995).

Drawing from the foundational work of Malle et al. (2014), but aiming to simultaneously manipulate all information processing elements, we devised a slightly simplified model that does not include agent's reasons for acting, nor agent's obligation to prevent the event. More precisely, we modified the causal model previously proposed by Franklin et al. (2022). Within our model (see Figure 1), we introduce the notion of *prevention effort*, causally connected to both foreseeability and capability. The model suggests that if people receive information about prevention efforts, they can immediately infer that the person has foreseeability and capability. By including prevention efforts in our model, we seek to examine to what extent this class of blame-relevant information can mitigate blame. Early research has indicated that people may be blamed more when they don't try to prevent bad outcomes (Knobe, 2003; Alicke et al., 2008).

The primary research questions are therefore: (1) Who is blamed more: *AVs* or human drivers? (2) Do people assess major classes of *blame-relevant information* differently for AVs and human drivers? (3) Is the novel element of *prevention effort* an influential class of information in people's blame judgements, both for human and AVs? (4) Do people judge other agents – beyond the AV or human driver – differently as a function of their causal connection to the AV or human driver (thus addressing the *many hands* issue)?

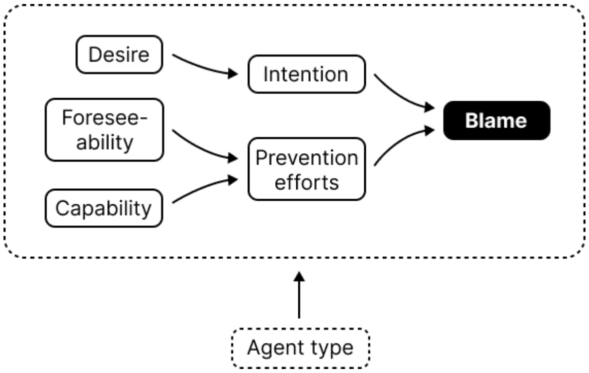

Figure 1: A proposed causal model of blame

Empirical research indicates that most people genuinely, not just metaphorically, assign blame to AI (Malle, Thapa, Scheutz, 2019; Stuart & Kneer, 2021) and that their blame judgements are influenced by their perception of the "cognitive" capacities of AI (Kneer & Stuart, 2021).

Empirical evidence on whether people assign different amounts of blame to AVs or humans is inconclusive. One study showed that people tend to blame AVs *less* than human drivers in accidents, such as pedestrian injuries caused by collisions between human-driven and autonomous vehicles (Li et al., 2016). By contrast, another study demonstrated that people tended to blame AVs *more* than human drivers, particularly when the severity of the outcome was more significant (Hong, Wang, & Lanz, 2020).

Zhang, Wallbridge, Jones, & Morgan (2021) studied six scenarios involving three types of outcomes (minor accidents, major accidents, or near-misses) caused by either human drivers or AVs. In five out of six scenarios that included an actual accident, AVs received more blame than humans. The sole scenario where human drivers were assigned more blame was a case where pedestrians crossing in front of a stopped bus were hit by a car passing the bus. The authors argued that this scenario presented more contextual cues and was more foreseeable and thus preventable. According to the authors, only a human driver could have understood the causal relationship between a stopped bus and the likelihood of nearby pedestrians. This argument suggests that the specific role of foreseeability and prevention efforts may differ in people's blame judgements of AVs as opposed to human drivers.

In the following experiment, we compare how people blame humans and AVs for their behaviours resulting in the same outcomes and examine in detail how they assess major classes of blame-relevant information. To explore the many hands problem in the case of AVs, we also probe alternative targets of blame, such as the company or programmer of an autonomous vehicle.

Method

Participants

We recruited 2647 international English-speaking participants (50% female, 37% students) through Prolific and compensated them £11 per hour. A total of 325 participants started the survey but did not finish it and were excluded from analyses, along with two participants who completed the survey hastily, evidently without reading the questions. The average age was 32 years (*SD* = 11), range 18-90. Most participants were from the United Kingdom (26%), South Africa (16%), Portugal (11%) and Poland (11%).

Design

The experimental design featured one between-subject factor of agent (a human taxi driver or an autonomous vehicle) and five between-subjects factors, two levels each, that represented the constructs from the causal blame model. These factors were (1) intentionality of bringing about the outcome (an accident), (2) desire to bring it about, (3) foreseeability of the outcome, (4) capability to prevent it, and (5) actual effort to prevent it. These five factors were manipulated to provide either positive evidence (e.g., the agent intentionally hit the other car) or negative evidence (e.g., the agent did not intentionally hit the other car). All participants saw information about all five factors, but which level they saw was randomly assigned. Thus, the overall design had 64 (2^6) cells. To account for the complexity of the design, we used a rule of thumb of an average of 40 participants per scenario to arrive at the sample of 2647.

Materials and measures

Participants read about a scenario that described the events leading up to a car crash. The scenario was, at least initially, purposefully inconclusive in order to make participants curious and desire more information. The events unfolded at a busy intersection, where a human-driven (but passengerless) taxi or a self-driving car collided with a second car, driven by the "victim", who was the only individual harmed by the crash. The extent of the injury was unspecified.

The main dependent variable was a blame judgement, measured on a scale from 0 (not at all) to 100 (completely), in response to the question, "How much do you blame the [taxi driver] / [self-driving car] for hitting the other car and injuring the person?" Participants engaged in two main rounds of blame assessment. In the first, they evaluated the focal agent, knowing only about the initial scenario description; and in the second round, they evaluated the agent after considering all information from the other five manipulated factors. Thereafter, participants also assigned blame to other agents (on a 0-100 scale). For the human taxi driver, these agents were the taxi company, the driver's instructor, and the victim; for the AV, the agents were the taxi company, the car's manufacturer, its programmer, and the victim of the accident.

We also assessed a number of other, exploratory variables, all on 0-100 rating scales. To account for how an agent's autonomy influences blame judgements, we asked participants: "To what extent did the agent behave autonomously?" We probed the objective foreseeability (Lagnado & Channon, 2008) of the crash, asking: "How likely was it that the agent would cause the crash?" Since our scenario implies causality without explicitly stating it, we further assessed participants' confidence in whether the agent or the other car caused the crash: (1) "How confident are you that the agent caused the crash?" and (2) "How confident are you that the other car caused the crash?" Finally, in the autonomous vehicle condition, we inquired into participants' anthropomorphism, asking "To what extent is the self-driving car human-like?"

Procedure

Participants were asked to assume the role of an investigator who had to decide how blameworthy the focal agent in the scenario (human driver or AV) was. After reading the initial scenario, participants made their first blame judgement, then received five different pieces of evidence as a text, one at a time (and each presenting one of the five experimental factors described above), and provided their final blame judgement with all information in mind. (We also assessed blame for each of the five pieces of evidence along the way, but because of space constraints we do not report these results.) Participants finally assigned blame to other agents, and responded to additional questions.

Results

We first tested whether people assigned blame differently to autonomous vehicles compared to human drivers. We then examined how blame-relevant information was used and tested the newly introduced prevention effort factor across agents. Finally, we inspected the role of other agents and analysed exploratory variables. T data is available at the Open Science Framework (OSF): osf.io/ts64j.

AV-Human Comparisons

After reading the initial scenario description, people blamed AVs more (M = 64.0, 95% CI [62.6, 65.4]) than human drivers (M = 59.8, 95% CI [58.4, 61.2]), $F_{(1, 2645)} = 17.07$, $p < .001$, $d = .16$ (see Figure 2, left panel). After consolidating all additional information they learned about the case, participants' final blame was higher ($M = 72.9$, $SD = 29.1$) than their initial one ($M = 61.9$, $SD = 26.1$), $F_{(1, 2646)} = 334.56$, $p < .001$, $d = .40$, but they no longer assigned different degrees of blame to the AV and human agent, $F_{(1, 2645)} < 1$, $p = .475$), as presented in Figure 2, right panel.

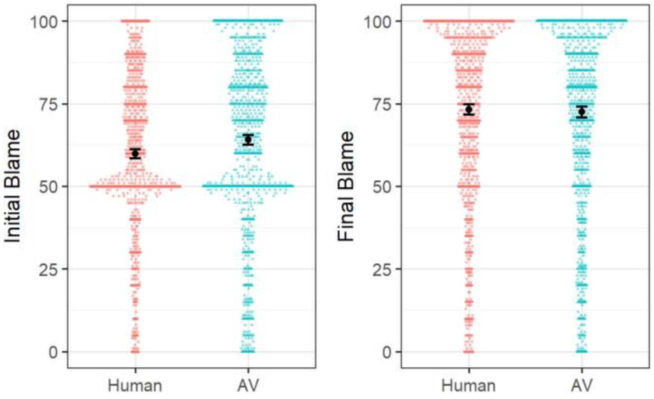

Figure 2: Distribution of initial (left) and final (right) blame by agent type

Next we examined which combinations of evidence predicted final blame judgements by running two separate analyses of covariance (ANCOVA). We first tested the intentional path (desire—yes/no × intentionality—yes/no), crossed with the agent factor (human/AV), with initial blame judgements serving as a covariate. Human agents received a slightly higher blame compared to AVs ($M_{\text{diff}} = 2.6$) when the initial blame was included as the covariate, $F_{(1, 2638)} = 6.31$, $p = .012$, $\eta_p^2 = .002$. Presence of intentionality increased blame by 8.9 points over absence of intentionality ($F_{(1,\ 2638)} = 75.56$, $p < .001$, $\eta_p^2 = .028$), and presence of desire increased blame by 5.5 points ($F_{(1,\ 2638)} = 28.87$, $p < .001$, $\eta_p^2 = .011$). In addition, a small interaction between intentionality and desire emerged ($F_{(1,\ 2638)} = 10.13$, $p = .001$, $\eta_p^2 = .004$), indicating that joint absence of the two factors or joint presence had slightly less impact on blame than presence of one in the absence of the other (see Figure 3, left panel). An interaction between intentionality and agent type ($F_{(1,\ 2638)} = 8.94$, $p = .003$, $\eta_p^2 = .001$) indicated that intentionality had a slightly stronger impact on blaming the human ($M_{\text{diff}} = -11.94$) than on blaming the AV ($M_{\text{diff}} = -5.83$). (See Figure 3, right panel).

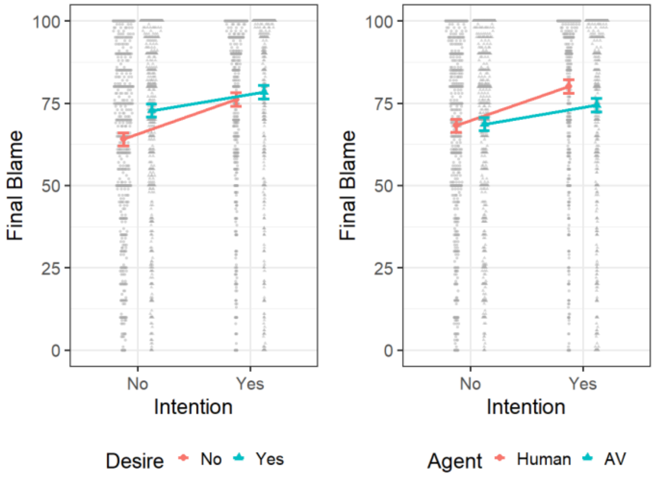

Figure 3: Final blame by intentionality and desire (left), and by intentionality and agent type (right)

Along the unintentional path, we conducted another ANCOVA, with capability, foreseeability, prevention effort (each with levels yes vs. no), and agent type, all as between-subjects factors, and initial blame as a covariate. Importantly, the analyses for the first time put the blame-guiding factor of *prevention effort* to the test. The human driver again received slightly more blame than the AV ($M_{\text{diff}} = 2.6$), $F_{(1,\ 2630)} = 6.50$, $p = .011$, $\eta_p^2 = .002$. We found an expected mitigation of blame when the crash was not foreseeable ($F_{(1,\ 2630)} = 53.88$, $p < .001$, $\eta_p^2 = .020$), or when the agent made efforts to prevent it ($F_{(1,\ 2630)} = 89.70$, $p < .001$, $\eta_p^2 = .033$). Against expectation, agents with low capability were blamed 4.6 points more than those with high capability ($F_{(1,\ 2630)} = 30.19$, $p < .001$, $\eta_p^2 = .011$). Two interactions emerged as well (see Figure 4). An interaction between foreseeability and prevention effort ($F_{(1,\ 2630)} = 9.74$, $p = .002$, $\eta_p^2 = .004$) indicated that the mitigating effect of prevention effort was stronger when the crash was not foreseeable than when it was foreseeable (akin to a little extra mitigation when the efforts were not expected). Finally, an interaction between capability and prevention effort ($F_{(1,\ 2630)} = 8.18$, $p = .004$, $\eta_p^2 = .003$) indicated that the mitigating effect of prevention effort was stronger when the agent was capable of preventing the crash than when the agent was not capable.

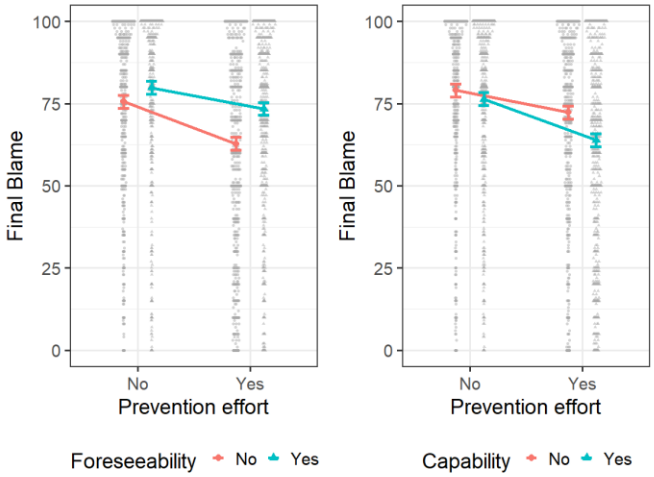

Figure 4: Final blame by prevention effort and foreseeability (left), and by prevention effort and capability (right)

Exploring the most impactful combinations of evidence

To generate future research hypotheses we inspected each of the combinations of 2^5 information conditions crossed with agent type and identified the highest and lowest blame judgements. (While Table 1 exclusively presents the highest and lowest mean final blame across all 64 groups, the full table is available at: osf.io/ts64j). The combination of factors that yielded the highest blame ($M = 91.5$, $SD = 16.4$) featured a human driver who had a desire to cause the crash, did it intentionally, had the capability to prevent it but did not exert effort to prevent the crash (surprisingly, the crash was described to be unforeseeable, but participants seemed to have dismissed this claim). The scenario with the second-highest blame rating ($M = 88.4$, $SD = 13.6$) involved a human driver who had a desire to cause the crash, in the end did not do it intentionally. The crash was foreseeable and the agent was capable of preventing it, yet did not make effort to do so.

Table 1: Selection of highest and lowest mean final blame

	Group[†]	M	Min	Max	SD	N
1	h11010	91.5	37	100	16.4	52
2	h01110	88.4	60	100	13.6	36
3	h11100	86.8	0	100	31.8	34
4	h11000	84.9	0	100	21.8	45
5	a01100	84.2	11	100	23.1	42
…	…	…	…	…	…	…
60	h00001	57.2	0	100	23.0	46
61	a00011	51.6	4	100	30.6	45
62	h00111	50.9	0	100	29.2	41
63	h01011	48.4	0	100	35.7	44
64	h00011	42.8	0	100	33.2	40

[†] In the Group column, the first symbol indicates the agent type (h = human, a = AV), followed by the factor levels (0 = no, 1 = yes) of intentionality, desire, foreseeability, capability, and prevention effort.

The scenario with the lowest blame rating (M = 42.8, SD = 33.2) involved a human driver who had no desire to cause a crash (which was unforeseeable), nor did they do it intentionally; further, they had the capability to prevent the crash and even tried to do so. The scenario with the second-lowest blame (M = 48.4, SD = 35.7) differed only in that the human driver had a desire to cause the crash, which was not enough for people to find the human agent especially blameworthy.

Among the first 14 groups with the highest blame ratings, a common thread emerged: Each scenario featured the agent failing to make an effort to prevent the crash. Conversely, instances of the lowest blame, for the most part, encompassed scenarios where the crash occurred unintentionally and yet the agent tried to prevent it. This indicates the crucial roles that both intentionality and the newly identified factor of preventive efforts play in shaping blame judgements.

Blaming Other Agents

To explore the extent of blame ascribed to other agents, we first conducted two one-way ANOVAs in which agent type (AV, human) was the independent variable. The first assessed blame for the associated company as the dependent variable, the second assessed blame for the accident victim.

There was a significant difference in blame judgements towards the company associated with the car, $F_{(1,\ 2645)}$ = 735.38, p < .001, d = 1.05. The company responsible for the AV received significantly higher blame (M = 68.4, 95% CI [66.7, 70.1]) than the company affiliated with the human taxi driver (M = 34.8, 95% CI [33.1, 36.5]).

Despite the fact that the victim was, on average, blamed less than the company, participants still found the victim to be more blameworthy in the scenario involving the AV compared to that with a human driver ($F_{(1,\ 2645)}$ = 30.74, p < .001, d = .22). The victim who crashed with an AV received significantly more blame (M = 26.3, 95% CI [24.9, 27.6]) than the victim involved in a collision with a taxi driver (M = 20.9, 95% CI [19.5, 22.2]).

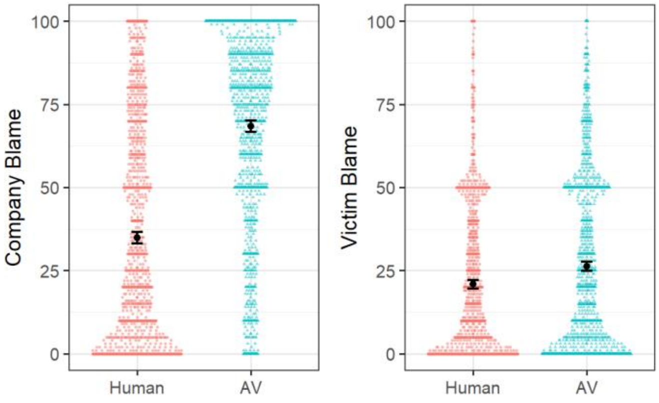

Figure 5: Company (left) and Victim (right) blame by agent

In a separate analysis, we explored the level of blame for the driving instructor in scenarios involving a human agent, comparing this with the level of blame for the programmer and manufacturer of an autonomous vehicle. Participants ascribed relatively low blame to the driving instructor, with a mean score of 21.1 (95% CI [19.5, 22.6]). By contrast, blame for the manufacturer of the autonomous vehicle was considerably higher, with a mean score of 65.7 (95% CI [64.0, 67.4]). Even more pronounced was the level of blame for the programmer of the AV, who received a mean blame rating of 70.9 (95% CI [69.4, 72.5]).

Confidence in Causal Contributions

Our analysis revealed that participants were slightly more confident that the human taxi driver caused the crash (M = 68.4) than that the AV caused the crash (M = 66.0), $F_{(1,\ 2645)}$ = 4.19, p = .041, d = .08. Conversely, participants were less confident that the other car (the victim) caused the crash when the primary agent was a human driver (M = 29.7) than when the primary agent was an AV (M = 36.3), $F_{(1,\ 2645)}$ = 33.99, p < .001, d = .23.

Control factors

Additional analyses suggested that observed differences in blame judgements cannot be ascribed to the autonomy of the agent, the objective foreseeability of the event, or anthropomorphic perceptions of the autonomous vehicle. We conducted three ANCOVAs with agent as one factor and autonomy/ objective foreseeability/ anthropomorphism as the second factor, with initial blame judgements serving as a covariate. In all three cases, we found no interaction with the agent. Ratings of autonomy for the human driver (M = 64.5) and the AV (M = 65.1) were indistinguishable. This suggests that participants viewed the decision-making capabilities of the human driver and the autonomous vehicle as similar, regardless of the inherently different nature of their operational mechanisms. Similarly, the rated objective foreseeability of the crash was similar for the human driver (M = 54.1) and the AV (M = 53.0), suggesting that participants did not perceive one type of agent as more predictably involved in incidents than the other. Moreover, participants rated the AV as relatively low in human likeness (M = 23.4, SD = 25.7). This finding suggests that the attribution of human-like characteristics to the AV was not a significant factor in the allocation of blame to the AV.

Discussion

Our results showed that early, inconclusive descriptions of events leading to a car crash elicited higher blame for autonomous vehicles than for human drivers. This result is in line with findings by Hong, Wang, and Lanz (2020) and Zhang et al. (2021), but it diverges from Li et al.'s (2016) result, where AVs received less blame than humans in scenarios involving pedestrian injury. In Hong et al.'s (2020) and Li et al.'s (2016) scenarios, there was a clear explanation available for what happened and *who was at fault*, unlike in our experiment where, at first, it was not clear what exactly happened and who was to blame. The discrepancy in results cannot be explained by differences in the level of detail in which the scenarios were described since our results do not differ from those of Hong et al. (2020).

It is important to note that we did not take into account the impact of outcome severity, which may serve as a mediating factor, as demonstrated in the previously mentioned studies. In our scenario, the outcome of the crash included a damaged car and an injured person. If the outcome severity was much lower (e.g., brushing a side of the other car without injuring the passenger in the other car), the human driver may receive more blame than the AV, as demonstrated by Zhang et al. (2021).

Despite the initially higher blame for the AV than for the human driver (when little was known about the scenario), final blame judgements did not differ substantially for AVs and humans. Along both intentional and unintentional paths of information processing, we did find slight differences in how much the agents increased from initial to final blame. Blame for the human driver increased slightly more (2.6 points) than blame for the AV increased, but the effect size was negligible in both cases.

When we examined the impact of various information factors on final blame, the patterns of impact were almost always the same for AV and human driver, except that evidence of intentionality influenced blame judgements more for human agents than for AVs. However, the effect size in this case, too, was negligible.

The proposed causal model served as a robust foundation for examining blame judgements across agents. Most well-known factors such as intentionality, desire, and foreseeability provided expected results—significantly higher blame if an agent had intentionality, had desire, and the event was foreseeable. The inclusion of the novel prevention effort factor proved to be particularly insightful: Our findings indicate that agents who demonstrated an effort to prevent the crash were blamed significantly less. This highlights the substantial impact of perceived preventative actions on blame judgements, suggesting that the assessment of an agent's attempt to avert an accident is a critical component in the evaluative process of assigning blame.

Our results included an unexpected finding regarding the impact of an agent's *capability* to prevent the negative outcome. Specifically, participants assigned more blame to agents introduced as less capable—even though previous theory and evidence suggested that agents exhibiting less capability cannot be expected to prevent the outcome and therefore are blamed less. This discrepant finding may stem from the way we operationalised capability. In Malle et al.'s (2014) framework, capability is defined as the cognitive or physical capacity to prevent a negative outcome. By contrast, in our study, we presented participants with information regarding *how good an agent was at driving*. This phrasing focuses on general driving skill, an ability, rather than on the capacity to prevent a particular crash. Therefore, participants tended to assign more blame to agents whose driving ability was described as low. This reinterpretation of the discrepancy is also compatible with our finding that agents who tried to prevent the crash were blamed less when they were "capable" than when they were not, perhaps because participants considered the agent's efforts to be more controlled and effective.

The analysis of specific combinations of information about the agent's desire, intentionality, and so forth, showed that agents who either want to and intentionally cause a crash or foresee it to happen and fail to prevent it, are blamed most strongly. By contrast, blame is low when agents unintentionally cause a crash even if they tried to avoid it. This highlights the important role of both intentionality and prevention effort as key factors in a causal chain of blame attribution.

It is important to note that the notion of blaming may have had distinct interpretations for humans and AVs. By manipulating different factors and the involvement of other agents, our goal was, in part, to better understand how people interpret and use blame in relation to human and autonomous agents. A dedicated study explicitly aimed at determining whether AVs are subjected solely to metaphorical blame could offer a clearer answer on this (e.g., by probing participants' reasoning). Although the proposed model of blame judgements may be adequate for human agents, it may not be sufficient to account for the complex task of assigning blame to autonomous vehicles. That is because people consider additional agents beside the AV when making sense of the AV's behaviours and outcomes. More precisely, other agents causally entangled with the AV's actions received significantly higher blame than corresponding agents entangled with a human driver—especially the company that owns the car and, disconcertingly, the human crash victim. The tendency to assign more blame to the victim in autonomous vehicle accidents arguably results from a tendency to look for sources of blame beyond the vehicle itself, and may be influenced by the belief in the superior capabilities of autonomous vehicles compared to human drivers. Hence, future research should consider controlling or measuring this variable. Additionally, the programmer and the manufacturer of an autonomous vehicle also received high blame. Thus, any future model that aims to predict people's blame judgements to autonomous vehicles must take into account the contributions of various agents beside the AV. Such an approach would better reflect the dynamics at play in situations where AVs are involved, and provide an opportunity to better tackle the problems of many hands and the responsibility gap. Ultimately, this should help us establish better legal guidelines for assigning blame and liability in cases where autonomous vehicles cause harm.

Acknowledgments

The authors express gratitude to the AI Objectives Institute for generously funding the data collection.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin, 126*(4), 556.

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin, 34*(10), 1371-1381.

Ashton, H. (2022a). Definitions of intent suitable for algorithms. *Artificial Intelligence and Law*, 1-32.

Ashton, H. (2022b). Defining and Identifying the Legal Culpability of Side Effects Using Causal Graphs. In *CEUR Workshop Proceedings* (Vol. 3087). CEUR Workshop Proceedings.

Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior, 90*, 363-371.

Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature News, 538*(7623), 20.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353-380.

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299-309.

Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. *AI & Society, 34*(1), 129-136.

De Jong, R. (2020). The retribution-gap and responsibility-loci related to robots and automated technologies: A reply to Nyholm. *Science and Engineering Ethics*, *26*(2), 727-735.

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.

Franklin, M., Ashton, H., Awad, E., & Lagnado, D. (2022, July). Causal Framework of Artificial Autonomous Agent Responsibility. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 276-284).

Gerstenberg, T., Ejova, A., & Lagnado, D. (2011). Blame the skilled. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Hong, J. W., Wang, Y., & Lanz, P. (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.

Kneer, M., & Stuart, M. T. (2021, March). Playing the blame game with robots. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 407-411).

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*(3), 190-194.

Konert, A., & Balcerzak, T. (2021). Military autonomous drones (UAVs)-from fantasy to reality. Legal and Ethical implications. *Transportation research procedia, 59*, 292-299.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754-770.

Li, J., Zhao, X., Cho, M. J., Ju, W., & Malle, B. F. (2016). *From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars* (No. 2016-01-0164). SAE Technical Paper.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology*, *33*(2), 101-121.

Malle, B. F., & Knobe, J. (2001). The distinction between desire and intention: A folk-conceptual analysis. Intentions and intentionality: *Foundations of social cognition, 45*, 67.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. *Robotics and well-being*, 111-133.

Monroe, A. E., & Malle, B. F. (2017). Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, *146*(1), 123.

Monroe, A. E., & Malle, B. F. (2019). People systematically update moral judgments of blame. *Journal of Personality and Social Psychology, 116*(2), 215.

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

Parlangeli, O., Currò, F., Palmitesta, P., & Guidi, S. (2023, September). Asymmetries in the Moral Judgements for Human Decision-Makers and Artificial Intelligence Systems (AI) Delegated to Make Legal Decisions. In *Proceedings of the European Conference on Cognitive Ergonomics 2023* (pp. 1-4).

Scheutz, M., & Malle, B. F. (2021). May machines take lives to save lives? Human perceptions of autonomous robots (with the capacity to kill). *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare*, 89-102.

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-27.

Van de Poel, I., & Nihlén Fahlquist, J. (2012). Risk and responsibility. In *Essentials of risk theory* (pp. 107-143). Dordrecht: Springer Netherlands.

Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in cognitive science*, *2*(3), 454-485.

Zhang, Q., Wallbridge, C. D., Jones, D. M., & Morgan, P. (2021, June). The blame game: Double standards apply to autonomous vehicle accidents. In *International Conference on Applied Human Factors and Ergonomics* (pp. 308-314). Cham: Springer International Publishing.