Journal of Experimental Social Psychology 111 (2024) 104582

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Case Report

Hazardous machinery: The assignment of agency and blame to robots versus non-autonomous machines☆

Rael J. Dawtry [a,*], Mitchell J. Callan [b]

[a] Department of Psychology, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom
[b] Department of Psychology, University of Bath, United Kingdom

ARTICLE INFO

Keywords:
Robots
Blame
Mind perception
Autonomy

ABSTRACT

Autonomous robots increasingly perform functions that are potentially hazardous and could cause injury to people (e.g., autonomous driving). When this happens, questions will arise regarding responsibility, although autonomy complicates this issue – insofar as robots seem to control their own behaviour, where would blame be assigned? Across three experiments, we examined whether robots involved in harm are assigned agency and, consequently, blamed. In Studies 1 and 2, people assigned more agency to machines involved in accidents when they were described as 'autonomous robots' (vs. 'machines'), and in turn, blamed them more, across a variety of contexts. In Study 2, robots and machines were assigned similar experience, and we found no evidence for a role of experience in blaming robots over machines. In Study 3, people assigned more agency and blame to a more (vs. less) sophisticated military robot involved in a civilian fatality. Humans who were responsible for robots' safe operation, however, were blamed similarly whether harms involved a robot (vs. machine; Study 1), or a more (vs. less; Study 3) sophisticated robot. These findings suggest that people spontaneously conceptualise robots' autonomy via humanlike agency, and consequently, consider them blameworthy agents.

Robots are performing increasingly complex tasks in various domains of human activity, such as logistics, healthcare, and hospitality (International Federation of Robotics, IFR Statistical Department, 2022a, 2022b; Schwab & Davis, 2018). Although they bring many benefits, such as relieving workers from 'dull, dirty, and dangerous work', some robots pose a hazard. Injuries involving robots have mainly occurred in industrial settings, but could become more widespread as robots are adopted for risky tasks such as driving, performing surgery, or fighting wars (Lin, Abney, & Bekey, 2011; Winfield et al., 2021).

Simultaneously, advances in artificial intelligence are equipping robots with greater *autonomy* - the ability to sense, plan and implement goal-directed behaviours, without external input, for an extended period (Bekey, 2005; Lin et al., 2011). When this behaviour causes injury, questions regarding responsibility will inevitably follow, yet autonomy complicates this issue. Insofar as the behaviour is – or at least seems to be – controlled by a robot itself, who should be held responsible? In three experiments, we investigated whether blame is assigned to robots involved in accidents. Specifically, we tested whether people blame targets labelled 'robots' more than 'machines', and whether they do so because robots are assigned humanlike agency.

1. Blaming robots

Blaming an unfeeling machine cannot avenge or deter future harm, and doing so thus seems pointless and irrational. Regardless of autonomy, robots' behaviour is determined by their programming and mechanical design, which depend upon human designers, users, and regulatory bodies. When injuries happen due to a design flaw, fault, or misuse of a machine, responsibility is typically assigned to one or more such parties. People seem to resolve responsibility for harms involving one type of robot, autonomous vehicles (AVs), in a similar way. Blame is directed toward parties that, whilst not immediately involved, are responsible for ensuring AV's safety, such as its manufacturer (Copp, Cabell, & Kemmelmeier, 2021; Li, Zhao, Cho, Ju, & Malle, 2016; McManus & Rutchick, 2019; Pöllänen, Read, Lane, Thompson, & Salmon, 2020).

☆ This paper has been recommended for acceptance by Professor. Rachel Barkan.

* Corresponding author.

E-mail addresses: rjdawt@essex.ac.uk (R.J. Dawtry), mjc213@bath.ac.uk (M.J. Callan).

https://doi.org/10.1016/j.jesp.2023.104582
Received 22 September 2023; Received in revised form 28 November 2023; Accepted 10 December 2023
Available online 19 December 2023

Insofar as robots at least *appear* to decide and control their behaviour, it has been suggested that people may nevertheless blame them (e.g., Bigman, Waytz, Alterovitz, & Gray, 2019; Malle, Magar, & Scheutz, 2019; Malle, Monroe, & Guglielmo, 2014). Li et al. (2016) found that, although autonomy increased blaming of a vehicle's manufacturer and government most, an AV was blamed more than a manual vehicle for an accident, although less than the manual's driver. Pöllänen et al. (2020) also found an AV was blamed versus a manual, although less than the AV's user/manual's driver. Studies using dilemma-like situations, in which an AV explicitly chooses a harm-causing action, have reported similar blame for an AV and its 'secondary driver' (Awad et al., 2018), and for a military drone and pilot (Malle et al., 2019), that chose similarly. Relatedly, robots are blamed more than humans for not making a utilitarian choice in a moral dilemma (Komatsu, Malle, & Scheutz, 2021; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015). These findings suggest that people can blame robots, even when other plausible targets are available, and especially when harm results explicitly from robots' autonomous choices.

AV accidents, however, may not be representative of those involving other robots and, although dilemmas reveal preferences regarding robots' moral behaviour, they have been criticised as unrealistic and unlikely (De Freitas, Anthony, Censi, & Alvarez, 2020; Roff, 2018). AI decision-making is opaque and technical, and robots cannot articulate their thoughts. Thus, whether and how harm is related to a robot's choices (e.g, an AV *choosing* to swerve; Awad et al., 2018) would be ambiguous much of the time (Chesterman, 2021; Langley, Meadows, Sridharan, & Choi, 2017). Relatedly, most actual robot injuries have occurred in industrial settings, and are ascribed to unsafe working practices, human error, or mechanical fault (Guiochet, Machin, & Waeselynck, 2017; Winfield et al., 2021). Such accidents occur as robots go about their routine, in a relatively 'mindless' way, and may appear very similar to those involving any kind of continuously operating machinery (e.g., an industrial loom) to ordinary observers.

Nevertheless, research shows that harm-causing agents are blamed relatively automatically (Alicke, 1992, 2000; Greene, 2007, 2009), even when the relevant behaviour is not of the agent's choosing (e.g., Woolfolk, Doris, & Darley, 2006). Accordingly, people may not need to know whether or how a robot's choices were involved - that it is the proximal cause of harm, and appears capable of autonomous behaviour *in general,* may be sufficient for people to blame a robot.

In the present studies, we tested whether people blame targets labelled 'autonomous robots' more than those labelled 'machines', for similar accidents. Besides how targets were labelled, accidents were described identically and we provided no information about targets' choices or internal states. Theorising on mind perception and responsibility suggests the effect of this manipulation should hinge on how robots' autonomy is *perceived*; specifically, the extent to which robots are assigned humanlike mental capacities, most crucially agency, that qualify the assignment of moral responsibility.

2. The role of agency and experience in robot blaming

Expertise is required to understand the material causes of robots' autonomous behaviour, and it has been suggested that, for most people, it is thus more a matter of perception than concrete facts (Bigman, Waytz, Alterovitz, & Gray, 2019). People anthropomorphise robots, relying on readily available knowledge of human mental capacities, such as agency, to explain how and why robots behave as they do. Agency encompasses mental abilities such as awareness of self, the environment and other beings, and the ability to plan, set goals, and volitionally enact different behaviours. Agency contrasts with *experience*, which encompasses emotions and sensation (Bigman & Gray, 2018; Gray, Gray, & Wegner, 2007; Gray & Wegner, 2012).

People readily assign agency to robots, albeit less than to humans (Bigman & Gray, 2018; Gray et al., 2007; Gray & Wegner, 2012; Yam et al., 2021), and seem to do so for two overarching reasons. First, robots move independently of external forces and often possess humanlike attributes that promote anthropomorphism (Epley, Waytz, & Cacioppo, 2007). Second, people are motivated to explain and predict other agents' behaviour, and assigning robots agency renders their behaviour comprehendible via readily available knowledge of human mental attributes (Waytz et al., 2010). Correspondingly, robots are assigned more agency when they behave unpredictably (Eyssel, Kuchenbrandt, & Bobinger, 2011), presumably in an effort to make sense of unexpected behaviour. Harmful behaviour by robots is unusual and may violate expectations - for example, that robots are precise – so could be especially prone to elicit agency attributions (cf. van der Woerdt & Haselager, 2019).

Ascribing agency to a robot may lead people to hold it responsible. Agency enables foresight of consequences, and the freedom to choose different behaviours in the same situation, and thus entails that a robot could have behaved differently than it did (e.g., a non-harmful way). Research suggests that possessing agency 'qualifies' an entity for responsibility, such that targets imbued with more agency are afforded greater moral responsibilities in general, and are blamed more for harming (Feinberg, Fang, Liu, & Peng, 2019; Gray et al., 2007; Gray & Wegner, 2009).

Whether experience is involved in blaming robots is less clear. People ascribe experience to robots under some conditions (e.g., when they appear humanlike; Gray & Wegner, 2009), but at relatively low levels (Gray et al., 2007). 'Dyadic morality' suggests that agency and experience are predominantly related to the affordance of, respectively, responsibility (for moral agents), and protection and concern (moral patients; Schein & Gray, 2018). Because consequences (e.g., retribution, deterrence) depend on the capacity to feel (e.g., remorse, suffering), however, assigning experience to a robot could make it seem a more valid target for blame, and thus enable people to blame it more readily.

3. The present research

Across three experiments with convenience samples of North American and UK internet-users, we examined attributions of agency and blame to robots, and tested whether people blame robots because they assign them agency. In Studies 1 and 2, participants read brief vignettes, involving various contexts and types of robot (e.g., healthcare, military, manufacturing), describing accidents in which a person was injured. We manipulated whether accidents involved an 'autonomous robot' or a 'machine'. Scenarios were otherwise identical between conditions and contained minimal information. We thus tested whether labelling a machine an 'autonomous robot' is sufficient to elicit agency and blame attributions. In Studies 1 and 2, we predicted more agency and blame would be assigned to robots (vs. machines), and that agency would mediate, such that robots (vs. machines) would be assigned more agency, and in turn, be blamed more. We did not test the alternative model, in which robots (vs. machines) are assigned greater agency *because* they are blamed more. Such a process seems unlikely given research showing that perceived mental states affect blame (for a review, see Malle, Guglielmo, & Monroe, 2014), rather than vice versa. Conceptually, there is no reason to expect robots to be blamed more than machines, for an identical harm, unless they are first perceived to have attributes (e.g., agency) that render them more blameworthy.

In Study 2, we also measured attributions of experience, and examined their contribution to blaming alongside agency. Study 3 sought corroborating evidence for the assumed causal process by which robot blaming occurs (Spencer, Zanna, & Fong, 2005); we varied information about objective capabilities mirroring subcomponents of agency, and predicted a robot would be blamed more when it seemed more (vs. less) agentic, supporting a causal role for agency in robot blaming. In Studies 1 and 3, we also explored blaming of humans who were responsible for robots' or machines' safe operation (e.g., health and safety manager).

We report sensitivity power analyses, measures, manipulations, and participant exclusions for all studies. An initial study (Supplementary

Study A), with results consistent with Studies 1 and 2, is reported in supplementary materials. All data, code, and materials are publicly accessible at: https://osf.io/npb9s/?view_only=1bf6445d96c242dbbe6346ac49229427

4. Study 1

4.1. Methods

4.1.1. Participants

One-hundred and forty-four participants were requested via Prolific Academic for a study on 'Perceptions of Accidents'. Of 147 who completed the survey, 2 failed at least one attention check, resulting in a final sample of 145 in the analyses below ($M_{age} = 38.09$, $SD_{age} = 13.20$, 94 females, 49 males, 2 other). This sample size gave 80% and 90% power to detect effects of target type on agency or blame attributions of $d_z = 0.23$ and 0.27, respectively.

4.1.2. Materials & procedure

Participants responded to brief vignettes, each describing an accident in which a person was injured by either an 'autonomous robot' or a 'machine', across 24 scenarios/settings. Each participant saw 8 (of 48 total) vignettes, 4 with a robot and 4 with a machine, in a randomized order. Every target was presented in a unique scenario for any participant. For example, in one scenario, participants read:

> *A worker at an automotive plant was seriously injured in an accident involving (an autonomous fabrication robot/a fabrication machine). The worker was making repairs to a conveyer belt in the (robot's/machine's) vicinity. They were impaled in the shoulder by a sharp metal rod the (robot/machine) was moving along the production line.*

Participants responded to six items about the target's agency adapted from Gray & Wegner, 2009, 2012, such as "The (robot/machine) is able to exert self-control" and "…can influence the outcome of situations". Three items asked about the target's blameworthiness; "The (robot/machine) caused this accident to happen", "… is morally responsible for this accident", and "… is blameworthy for this accident". Finally, participants judged the blameworthiness of a (human) manager, who was described as "… directly responsible for ensuring employee's health and safety", using the same three blame items. All responses were on 7-point scales (1 = *Strongly Disagree*; 7 = *Strongly Agree*). Agency (α = 0.86), blame (α = 0.85), and manager blame (α = 0.89) measures had good reliability, and items were averaged together within each. Participants also answered simple attention checks ('… select 'strongly agree/disagree") embedded in measures following two vignettes.

4.2. Results

Agency, target blame, and manager blame attributions were fit in separate mixed effects models, each including a fixed effect of target type, random intercepts for participants and scenarios, and random slopes for the effects of target type by participants and by scenarios, using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015, Version 1.1–33) in R (R Core Team, 2021, Version 4.3.0). Random effects were correlated. We used Satterthwaite approximations to calculate p values using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017, Version 3.1–3), and report 95% percentile bootstrap confidence intervals (CIs; 2000 resamples).

As shown in Fig. 1, these analyses revealed significant effects of target type on both agency and blame attributions. Robots were assigned more agency, $b = 0.27$, $SE = 0.05$, 95% CI [0.16, 0.37], $t(46.57) = 5.17$, $p < .001$, and blamed more, $b = 0.29$, $SE = 0.08$, 95% CI [0.13, 0.43], $t(21.10) = 3.78$, $p = .001$, than machines. Attributions of blame to managers, however, were similar whether the target was labelled a 'robot' or a 'machine', $b = 0.03$, $SE = 0.06$, 95% CI's [−0.08, 0.15], $t(122.78) = 0.57$, $p = .57$.

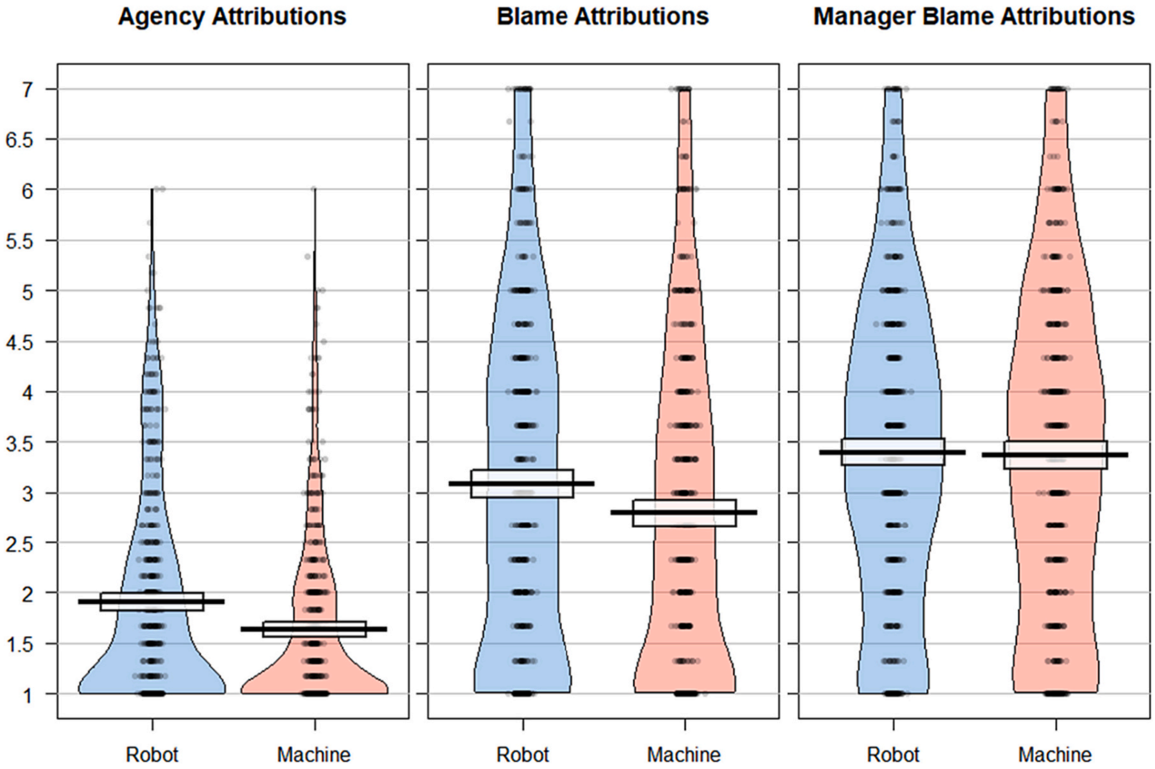

Fig. 1. Raw data, descriptive and inferential statistics plots (using the "yarrr" package in R; Phillips, 2017, Version 0.1.5) of the effects of target type on agency, blame, and manager blame attributions. The black horizontal lines show means within conditions, and the error bars are 95% CIs.

Agency and blame attributions were positively related for both machine, *r* = .41, *p* < .001, and robot targets, *r* = .40, *p* < .001. We next employed the approach of joint significance testing advocated by Yzerbyt, Muller, Batailler, and Judd (2018) to test the mediated effect of target type on blame through agency attributions. This approach assumes that, if the effect of target label on agency is statistically significant (a path), and the direct effect of agency on blame is statistically significant (b path), it follows that the mediated effect of target type on blame is significantly different from zero. Analyses were performed using the lme4 package in R as above. Robot vs. machine was coded 0.5 vs. -0.5. Estimated path coefficients are shown in Table 1. The effect of robot vs. machine on agency attributions (a path) was statistically significant. To test the direct relationship between agency and blame (b path), we fit a model predicting blame attributions that included a robot vs. machine contrast and agency attributions as fixed effects. We included random intercepts for participants and scenarios, and random slopes by participants and by scenarios for the effect of target type and the relationship between agency and blame. Analyses revealed a significant direct relationship between agency and blame; thus, in conjunction with the significant a path, evidencing an indirect effect of robot vs. machine on blame through agency attributions, shown in Fig. 2 (top). We also tested the indirect effect of target type on blame through agency using the quasi-Bayesian Monte Carlo method (5000 simulations) with the *mediation* package (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014; see Imai, Keele, & Tingley, 2010) in R. These analyses revealed significant mediated effects when performed grouping only by participants (indirect effect = 0.15, 95% CI [0.08, 0.23]; *p* < .001) and only by scenarios (indirect effect = 0.17, 95% CI [0.06, 0.29]; *p* = .002) (note that the software only allows for one group type per model). This suggests that the indirect effect of target type on blame through agency generalizes across participants and across scenarios.

In sum, agency and blame were positively related, and similarly so for 'autonomous robots' and 'machines' separately. Because robots were assigned more agency, however, they were, in turn, blamed more. This effect occurred across a range of settings, involving robots serving a variety of functions (e.g., surgery, cleaning). The type of machine, however, had no impact on blaming of a human target with responsibility for preventing harm; the manager was blamed similarly regardless of whether accidents involved a robot or a machine. In a subsequent study, we sought replicate these findings and additionally examined the role of experience in robot blaming.

Table 1
Estimated coefficients for the mediated effects of target type on blame attributions through agency attributions.

Outcome:		Agency Attributions	Blame Attributions		
		B [95% CI]	*B* [95% CI]		*B* [95% CI]
Study 1					
Robot vs. Machine	*a*	0.27* [0.16, 0.37]	0.29* [0.13, 0.43]	*c'*	0.14 [−0.0, 0.28]
Agency				*b*	0.57* [0.45, 0.70]
Study 2					
Robot vs. Machine	*a*	0.28* [0.19, 0.38]	0.34* [0.21, 0.47]	*c'*	0.17* [0.05, 0.28]
Agency				*b*	0.54* [0.36, 0.71]
Experience		1.00* [0.79, 1.20]	0.71* [0.37, 1.05]		0.05 [−0.27, 0.39]

Note. a = effect of target on agency attributions. c = total effect of target on blame attributions. c′ = the direct effect of target on blame attributions. b = the direct relationship between agency and blame attributions.

* $p < .05$.

5. Study 2

5.1. Methods

5.1.1. Participants

One-hundred and forty-four participants were requested via Prolific Academic for a study on 'Perceptions of Accidents'. We sought to replace 10 participants who failed at least one attention check, resulting in a final sample of 143 in the analyses below ($M_{\text{age}} = 40.08$, $SD_{\text{age}} = 12.89$, 97 females, 46 males). This sample size gave 80% and 90% power to detect effects of target type on agency, experience or blame attributions of $d_z = 0.24$ and 0.27, respectively.

5.1.2. Materials & procedure

Study 2 was identical to Study 1, except for the addition of four items on targets' experiential mental capacities, adapted from Gray & Wegner, 2009, 2012, such as "The (robot/machine) can experience emotions". Also, measures of managers' blameworthiness were not included. The agency (α = 0.84), experience (α = 0.98), and blame (α = 0.85) measures had good reliability, and items were averaged together in each.

5.2. Results

Linear mixed effects analyses as per Study 1 revealed significant effects of target type on agency and blame attributions: as shown in Fig. 3, robots were assigned more agency, *b* = 0.30, *SE* = 0.05, 95% CI [0.19, 0.40], *t*(83.94) = 5.67, *p* < .001, and blamed more, *b* = 0.35, *SE* = 0.07, 95% CI [0.22, 0.48], *t*(59.71) = 4.93, *p* < .001, than machines. Attributions of experience, however, were similar whether the harm-doing target was labelled a robot or machine, *b* = 0.01, *SE* = 0.01, 95% CI's [−0.02, 0.04], *t*(31.95) = 0.78, *p* = .44.

Agency and blame attributions were positively related for both machine, *r* = .29, *p* < .001, and robot targets, *r* = .37, *p* < .001. Experience attributions were positively related to blame for machine, *r* = .17, *p* = .039, but not robot targets, *r* = .11, *p* = .197. Similarly to Study 1, we tested the mediated effect of target type on blame through agency attributions while controlling for experience attributions, using the approach of joint significance testing. Robot vs. machine was coded 0.5 vs. -0.5; estimated path coefficients are shown in Table 1. The effect of robot vs. machine on agency (a path) while controlling for experience was statistically significant (this model included fixed effects for target and experience and by-participant and by-scenario random intercepts and random slopes for the effect of target and the relationship between experience and agency). To test the direct relationship between agency and blame attributions (b path) while controlling for experience attributions, we fit a model predicting blame that included a robot vs. machine contrast, agency, and experience attributions as fixed effects. The model also included random intercepts for participants and scenarios, and random slopes by participants and by scenarios for the effects of target type, agency, and experience. Analyses revealed a statistically significant direct relationship between agency and blame; thus, in conjunction with the significant a path, evidencing an indirect effect of robot vs. machine on blame through agency attributions while controlling for experience, shown in Fig. 2 (bottom). Tests of the indirect effects per Study 1 revealed significant mediated effects when analyses were performed grouping only by participants (indirect effect = 0.17, 95% CI [0.09, 0.26]; *p* < .001) and only by scenarios (indirect effect = 0.20, 95% CI [0.10, 0.32]; *p* < .001). This shows that the indirect effect of target type on blame through agency while controlling for experience generalizes across participants and across scenarios.

Study 2 replicated the Study 1 findings: agency was positively related to blaming of targets labelled 'autonomous robots' and 'machines' separately, although because more agency was assigned to robots than machines, robots were, in turn, blamed more. We found no evidence that experience contributed to greater blaming of robots than machines: robot and machine targets were assigned similar, near-zero levels of

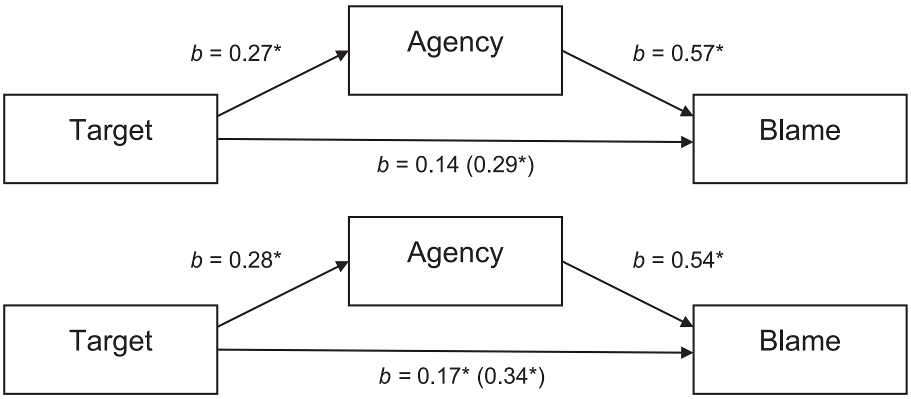

Fig. 2. Path models of the indirect effect of target (robot vs. machine) on blame via agency attributions in Studies 1 (top) and 2 (bottom).
Note. Robot and machine targets were coded 0.5 and − 0.5, respectively. Total effects are shown in parentheses. Estimated path coefficients for Study 2 (bottom) control for experience attributions.
* $p < .05$.

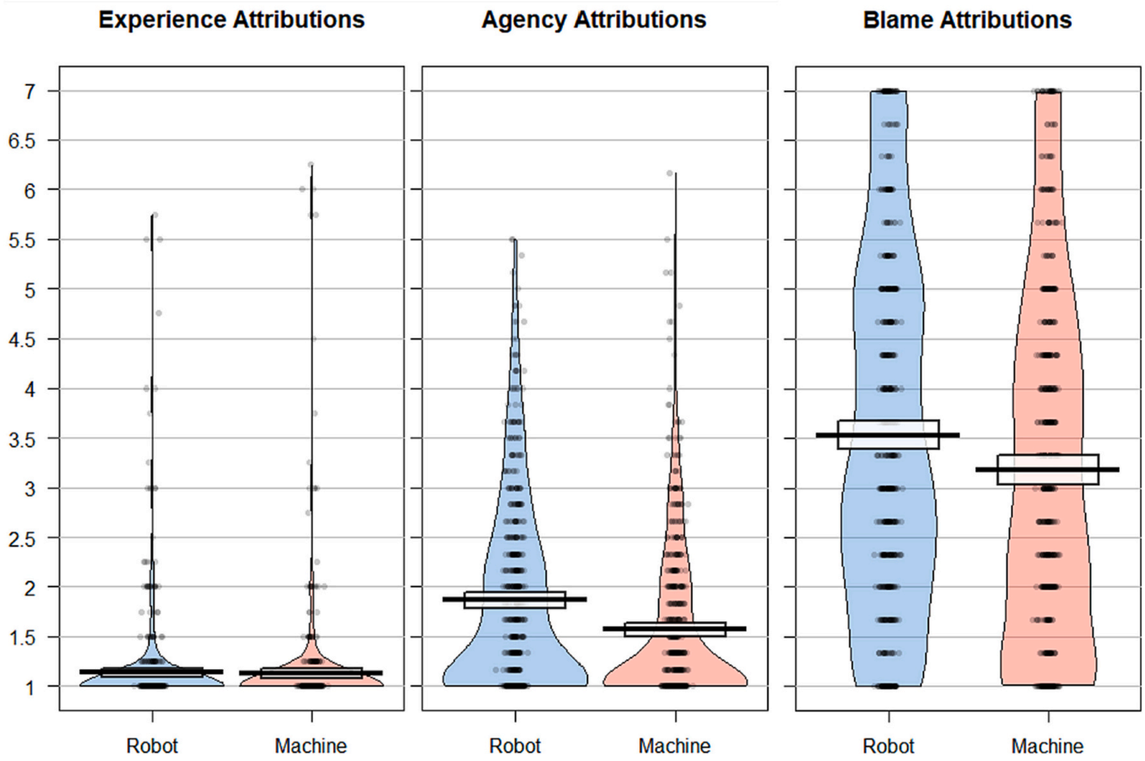

Fig. 3. Raw data, descriptive and inferential statistics plots (using the “yarrr” package in R; Phillips, 2017, Version 0.1.5) of the effects of target type on experience, agency, and blame attributions. The black horizontal lines show the means within conditions, and the error bars are 95% CIs.

experience, and agency mediated the effect of target type on blaming over and above experience. In Study 3, we sought corroborating evidence for the assumed causal process by directly manipulating perceptions of a robot's agency, and measuring blame. We also explored whether robots' level of agency affected blaming of human targets who were responsible for the robot.

6. Study 3

6.1. Method

6.1.1. Participants

Two-hundred and eighty participants were requested via Prolific Academic for a study on ‘Perceptions of Robots’. We replaced participants (n = 52) who failed one or more attention checks, resulting in a final sample of 282 in the analyses below (M_{age} = 38.9, SD_{age} = 13.41, 178 females, 100 males, 3 other, 1 unreported). This sample size gave 80% and 90% power to detect effects of target type on attributions of d = 0.33 and 0.39, respectively.

Table 2
Descriptive statistics for Study 3.

	Robot's level of sophistication				
	Less	More			
Measure	*M (SD)*	*M (SD)*	*1*	*2*	*3*
1. Robot's agency	3.15 (1.13)	4.24 (1.05)	–		
2. Blame (Robot)	2.95 (1.71)	3.53 (1.81)	0.39*	–	
3. Blame (Officer)	5.03 (1.53)	5.02 (1.65)	−0.11	−0.16*	–
4. Blame (Operator)	4.90 (1.61)	4.73 (1.65)	−0.14*	0.003	0.63*

Note. *$p < .05$. $df = 280$ for all correlations.

6.1.2. Materials & procedure

Participants read a description of an 'armed bi-pedal/humanoid tactical robot' - ABTAC - described as having more versus less sophisticated capabilities across two between-subjects experimental conditions.[1] Specifically, we varied information about capabilities mirroring subcomponents of agency, such as sensors and person/object recognition (awareness of self/others/environment), navigation and target tracking (planning), and level of oversight from a human (volition). We expected the more (vs. less) sophisticated robot would be assigned higher agency.

Participants next rated the robot's agentic mental capacities per items in Studies 1 and 2, which had good reliability and were averaged together (α = 0.81). An attention check ('… select 'strongly agree") was embedded in these items. Participants then read a scenario describing a military raid on a compound to capture or kill a terrorist. In the scenario, a commanding officer ordered a 'technical operator' to deploy the robot. Under circumstances described as '….not fully clear', the robot discharged its machinegun, and on entering the compound, the team discovered two deceased men, and an injured civilian girl who later died.

On the following page, participants responded to an attention check in which they identified which of four statements about the scenario was true ('The terrorist was captured', 'Three men were killed', 'A teenage girl later died from her wounds', 'Firearms were discovered at the compound'). Finally, participants assigned blame to the robot ($r = .44$, $p < .001$), commanding officer ($r = .79$, $p < .001$), and technical operator ($r = .80$, $p < .001$) for what happened to the teenage girl, per the 'moral responsibility' and 'blameworthiness' items in Studies 1 & 2, which were averaged within each target.

6.2. Results

Means by condition and intercorrelations among the Study 3 measures are shown in Table 2. The more sophisticated robot was assigned higher agency than the less sophisticated robot $t(275.25) = 8.36$, $p < .001$, $d = 0.99$. A 2 (robot's sophistication: lower vs. higher) *3 (target: robot, officer, operator) ANOVA, with repeated measures on the second factor, was conducted on blame attributions. There was no main effect of the robot's sophistication on blame, $F(1, 280) = 1.09$, $p = .30$, $\eta_p^2 = 0.004$. The main effect of target, $F(1.47, 412.73) = 114.56$, $p < .001$, $\eta_p^2 = 0.29$, and the sophistication by target interaction, were both significant $F(1.47, 412.73) = 4.75$, $p = .017$, $\eta_p^2 = 0.02$ (degrees of freedom were Greenhouse-Geisser corrected). As shown in Fig. 4, the more (vs. less) sophisticated robot was blamed more, $t(280) = 2.80$, $p = .005$, $d = 0.33$, although the robots' sophistication had no effect on blaming of the officer, $t(279.70) = 0.05$, $p = .96$, $d = 0.01$, or operator, $t(79.69) = 0.88$, $p = .38$, $d = 0.10$.

7. General discussion

Research investigating moral judgments of robots has often relied on dilemma-like scenarios that overtly implicate robots' decisions in causing harm. Consequently, it is unclear whether under ambiguous – and arguably more realistic – circumstances, observers will blame robots, and *more so* than they would any other machine. We addressed these issues by asking participants to assign blame for accidents involving targets labelled either 'autonomous robots' or 'machines', predicting that, insofar as people assign higher agency to robots than machines, robots would be blamed more.

Correspondingly, in Studies 1 and 2, people assigned more agency and blame to robots than machines. Agency mediated effects of target type in both studies, suggesting that robot blaming occurs *because* robots are assigned more agency which, in turn, renders them relatively blameworthy. Study 3 corroborated this assumed causal process; when agency was manipulated by varying information about a robots' sophistication, a more (vs. less) agentic-seeming robot was blamed more.

Simply labelling machines 'autonomous robots' (vs. 'machines') thus increased agency and blame, even though targets were otherwise described identically. Because all robots are machines, and some machines are robots, targets within matched scenarios *could well be identical* in actuality (e.g., similarly autonomous). Potentially, labels had a stereotype-like effect; 'autonomous robots' may form a category distinct from 'machines', which is automatically assumed to possess some humanlike qualities, including agency. This echoes findings on dehumanisation and stereotyping; for example, in-groups are attributed more humanlike minds than out-groups (for a review, see Kteily & Landry, 2022).

In Study 2, robots and machines were assigned similar, near-zero levels of experience; experience did not mediate the effect of target type on blame, and when agency and experience were examined simultaneously, only agency significantly predicted blame. People blamed robots, which were perceived as agentic but unfeeling, more than machines, which were perceived to lack agency *and* feelings, regardless of the apparent futility of blaming *any* machine. This is consistent with dyadic morality, which suggests that responsibility judgments depend mainly on perceived agency, rather than experience (Gray et al., 2007; Schein & Gray, 2018). Study 2 suggests that experience is not necessary for robots to be blamed, and does not explain why they are blamed more than machines. Because experience was at floor, however, and did not vary across machine and robot targets, it may nevertheless facilitate blaming robots under other conditions, when it does vary between or within targets.

It has been theorized that autonomy could displace responsibility, such that blame will be differently apportioned or overall reduced for harms involving autonomous machines (Matthias, 2004; Sparrow, 2007). Research on AV's, for example, finds that autonomy reduces blaming of a vehicle's 'driver', whilst increasing blaming of parties responsible for ensuring safety, such as an AV's manufacturer (e.g., Li et al., 2016; McManus & Rutchick, 2019). We explored whether indirectly responsible humans, such as a health and safety manager, were differently blamed when accidents involved robots (vs. machines; Study 2), and when a robot seemed more (vs. less; Study 3) agentic, but no differences emerged.

Who should be held accountable when robots harm people is important in ethical and legal debates, for example around autonomous weapons (Krishnan, 2009). So-called 'killer robots' could complicate the assignment of moral and legal responsibility, and thus dilute humanitarian considerations in the prosecution of war (Sparrow, 2007). In

[1] For example, participants read that ABTAC is equipped with *'state of the art visual, audio, and haptic (touch-sensing) sensory technology, which affords the robot a detailed and accurate representation of its environment, and its own location and interactions with the environment.'* (More sophisticated) versus '*cameras, microphones and motion-sensors which allows the robot to monitor and record its environment, and represent its own location within it.'* (Less sophisticated). This manipulation was verified beforehand in a separate study ($N = 45$, $M_{age} = 36.7$, $SD_{age} = 13.76$, 28 females, 17 males); the robot was assigned higher agency when it was more ($M = 4.41$, $SD = 0.89$) versus less sophisticated ($M = 3.38$, $SD = 1.07$), Welch's $t(36.89) = 3.42$, $p = .002$, $d = 1.05$.

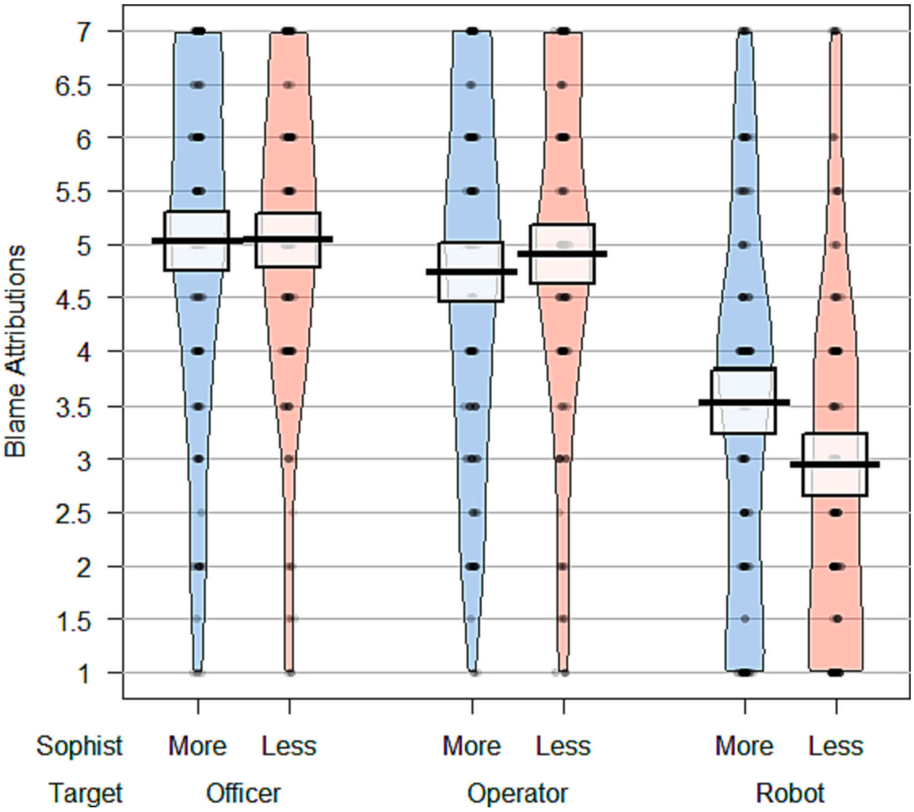

Fig. 4. Effect of robot's sophistication on blaming of the robot, officer, and operator.

Study 3, we found that a sophisticated (vs. less sophisticated) military robot was blamed for a civilians' death suggesting that, via agency attributions, autonomy does impact how people assign responsibility in such settings. Echoing findings on AV's, however, regardless of whether accidents involved robots or machines (Study 1, Supplementary Study A), or robots' level of sophistication (Study 3), human targets were always blamed substantially more than robots.

Although we found no evidence that autonomy impacts responsibility of humans who are responsible for robots, this issue warrants further scrutiny. Future research should examine factors that could modulate blame distribution across robots and humans (e.g., robots' appearance), and examine a broader range of targets (e.g., across the chain of command in military settings, robots' programmers; cf. Malle et al., 2019). Because attitudes and perceptions of robots vary across cultures (e.g., Bartneck, Suzuki, Kanda, & Nomura, 2007), it is also important to establish the generalisability of these findings across different samples.

8. Conclusion

As technology advances, robots are set to become more sophisticated and autonomous, thus expanding their uses. Some of the functions robots will, or already do, perform - on roads, factories, or the battlefield – are inherently hazardous, and will sometimes lead to injury, and in turn, attempts to determine responsibility. The present results suggest that, in such circumstances, insofar as robots are imbued with agency – and presumably, robots will appear more agentic as their autonomy grows – people may assign some responsibility to the robot itself.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data, code, and materials are available at https://osf.io/npb9s/?view_only=1bf6445d96c242dbbe6346ac49229427

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2023.104582.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*, 368. https://doi.org/10.1037/0022-3514.63.3.368

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*, 556. https://doi.org/10.1037/0033-2909.126.4.556

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., … Rahwan, I. (2018). Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation. *PsyArXiv.*. https://doi.org/10.48550/arXiv.1803.07170

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2007). The influence of people's culture and prior experiences with Aibo on their attitude towards robots. *AI & SOCIETY, 21*, 217–230. https://doi.org/10.1007/s00146-006-0052-7

Bates, D., Maechler, M., Bolker, B., & Walker, S. G. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Bekey, G. A. (2005). *Autonomous robots: From biological inspiration to implementation and control.* MIT press.

Bigman, E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*, 365–368. https://doi.org/10.1016/j.tics.2019.02.008

Bigman, Y., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Chesterman, S. (2021). Through a glass, darkly: Artificial intelligence and the problem of opacity. *The American Journal of Comparative Law, 69*, 271–294. https://doi.org/10.1016/j.tics.2019.02.008

Copp, C. J., Cabell, J. J., & Kemmelmeier, M. (2021). Plenty of blame to go around: Attributions of responsibility in a fatal autonomous vehicle accident. *Current Psychology, 42*, 6752–6767. https://doi.org/10.1007/s12144-021-01956-5

De Freitas, J., Anthony, S. E., Censi, A., & Alvarez, G. A. (2020). Doubting driverless dilemmas. *Perspectives on Psychological Science, 15*, 1284–1288. https://doi.org/10.1177/1745691620922201

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review, 114*, 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., Kuchenbrandt, D., & Bobinger, S. (2011). Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of the 6th international conference on Human-robot interaction* (pp. 61–68). https://dl.acm.org/doi/abs/10.1145/1957656.1957673.

Feinberg, M., Fang, R., Liu, S., & Peng, K. (2019). A world of blame to go around: Cross-cultural determinants of responsibility and punishment judgments. *Personality and Social Psychology Bulletin, 45*, 634–651. https://doi.org/10.1177/0146167218794631

Gray, H. M., Gray, K., & Wegner, D. M. (2007). *Dimensions of mind perception. science, 315*, 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*, 505–520. https://doi.org/10.1037/a0013748

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*, 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences, 11*, 322–323. https://doi.org/10.1016/j.tics.2007.06.004

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology, 45*, 581–584. https://doi.org/10.1016/j.jesp.2009.01.003

Guiochet, J., Machin, M., & Waeselynck, H. (2017). Safety-critical advanced robots: A survey. *Robotics and Autonomous Systems, 94*, 43–52. https://doi.org/10.1016/j.robot.2017.04.004

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309–334.

International Federation of Robotics, IFR Statistical Department. (2022a). *World Robotics 2022 - Industrial Robots*. Retrieved from: https://ifr.org/img/worldrobotics/Executive_Summary_WR_Industrial_Robots_2022.pdf.

International Federation of Robotics, IFR Statistical Department. (2022b). *World Robotics 2022 - Service Robots*. Retrieved from: https://ifr.org/img/worldrobotics/Executive_Summary_WR_Industrial_Robots_2022.pdf.

Komatsu, T., Malle, B. F., & Scheutz, M. (2021, March). Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across US and Japan. In *In Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 63–72). https://doi.org/10.1145/3434073.3444672

Krishnan, A. (2009). *Killer robots: legality and ethicality of autonomous weapons*. London: Routledge.

Kteily, N. S., & Landry, A. P. (2022). Dehumanization: Trends, insights, and challenges. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2021.12.003

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26.

Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable agency for intelligent autonomous systems. *In Proceedings of the AAAI Conference on Artificial Intelligence, 31*(2), 4762–4763. https://doi.org/10.1609/aaai.v31i2.19108

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. *SAE Technical Papers, 10*. https://doi.org/10.4271/2016-01-0164, 2016–01.

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence, 175*, 942–949. https://doi.org/10.1016/j.artint.2010.11.026

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. In M. I. A. Ferreira, J. S. Sequeira, G. S. Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (pp. 111–133). https://doi.org/10.1007/978-3-030-12524-0_11

Malle, B. F., Monroe, A. E., & Guglielmo, S. (2014). Paths to blame and paths to convergence. *Psychological Inquiry, 25*, 251–260. https://doi.org/10.1080/1047840X.2014.913379

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. In *, 117-124. In Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction.* https://doi.org/10.1145/2696454.2696458

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183. https://doi.org/10.1007/s10676-004-3422-1

McManus, R. M., & Rutchick, A. M. (2019). Autonomous vehicles and the attribution of moral responsibility. *Social Psychological and Personality Science, 10*, 345–352. https://doi.org/10.1177/1948550618755875

Phillips, N. (2017). *yarrr: A companion to the e-book "YaRrr!: The Pirate's Guide to R". R package version 0.1.5*. Retrieved from https://CRAN.R-project.org/package=yarrr.

Pöllänen, E., Read, G. J., Lane, B. R., Thompson, J., & Salmon, P. M. (2020). Who is to blame for crashes involving autonomous vehicles? Exploring blame attribution across the road transport system. *Ergonomics, 63*, 525–537. https://doi.org/10.1080/00140139.2020.1744064

R Core Team. (2021). *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org.

Roff, H. (2018). *The folly of trolleys: Ethical challenges and autonomous vehicles*. The Brookings Institution. https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Schwab, K., & Davis, N. (2018). *Shaping the fourth industrial revolution*. Geneva: World Economic Forum.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*, 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*(6), 845–851. https://doi.org/10.1037/0022-3514.89.6.845

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014a). Mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59*, 1–38. http://www.jstatsoft.org/v59/i05/.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J. H., & Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology, 99*, 410. https://psycnet.apa.org/doi/10.1037/a0020240.

Winfield, A. F., Winkle, K., Webb, H., Lyngs, U., Jirotka, M., & Macrae, C. (2021). Robot accident investigation: A case study in responsible robotics. In A. Cavalcanti, B. Dongol, R. Hierons, J. Timmis, & J. Woodcock (Eds.), *Software engineering for robotics* (pp. 165–187). https://doi.org/10.1007/978-3-030-66494-7_6

van der Woerdt, S., & Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology, 54*, 93–100. https://doi.org/10.1016/j.newideapsych.2017.11.001

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*, 283–301. https://doi.org/10.1016/j.cognition.2005.05.002

Yam, K. C., Bigman, Y. E., Tang, P. M., Ilies, R., De Cremer, D., Soh, H., & Gray, K. (2021). Robots at work: People prefer—And forgive—Service robots with perceived feelings. *Journal of Applied Psychology, 106*, 1557.

Yzerbyt, V., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology, 115*(6), 929–943. https://psycnet.apa.org/doi/10.1037/apl0000834.