In the loop, out of sync: Moral judgment in human-AI control architectures

Ritsaart Reimann (ritsaart.reimann@uni-graz.at) & Markus Kneer (markus.kneer@uni-graz.at)

University of Graz

Abstract

The ubiquitous integration of AI-powered systems in morally consequential decision-making procedures raises a thorny question: when such systems generate harm, who should be held responsible? A prominent regulatory response proposes that suitably designed control architectures can ensure that blame is appropriately allocated. To live up to this promise, the proposed architectures must be both normatively adequate *and* regarded as such, since otherwise they are at best practically useless, and at worst useless *because* morally problematic. We examine two of the most widely discussed proposals. Experiment 1 (n=260) investigates whether laypeople attribute responsibility differently across architectures that place human agents in-, on-, and out-of-the-loop. Experiment 2 (n=520) extends this inquiry by differentiating between control structures that either violate or satisfy Santoni de Sio's and van den Hoven's (2018) 'track and trace' requirements. Our results reveal that neither proposal fully succeeds in directing laypeople's responsibility judgements along the pathways they prescribe. Implications are discussed.

Keywords: responsibility attribution; control architectures.

Introduction

The ubiquitous integration of AI-powered systems in morally consequential decision-making procedures has been shadowed by an expanding catalogue of harms (e.g., Citron & Pasquale, 2014; Barocas & Selbst, 2016; Eubanks, 2018). A hotly contested question that arises in this context concerns the attribution of blame, given that it is often unclear who, if anyone, can be justly held responsible for the actions of autonomous systems (Nyholm, 2022). Who, for instance, should be held accountable when a self-driving car veers off the road and injures an innocent bystander? According to a number of recent responses, these sorts of cases may be less intractable than they initially appear (e.g., Nyholm, 2018; Coeckelbergh, 2019; Tigard, 2021). Köhler and colleagues (2018, see also Königs, 2022), for instance, argue that familiar standards of negligence and recklessness cover and in that sense 'plug' the vast majority of so-called 'responsibility-gaps', i.e., situations where some harm occurs, but where no human agent can reasonably be deemed responsible for that harm. Others insist that such gaps are both genuine and pervasive, arguing that the non-deterministic nature of contemporary learning automata renders their outputs *in principle* impossible to predict, and that this unpredictability can make it impossible to *justly* attribute blame (e.g., Matthias, 2004; Sparrow, 2007). We call this conviction the *No Blame Principle*.

Amongst countless gaps that have since been diagnosed (for a review, see Santoni de Sio & Mecacci, 2021; Oimann & Tollon, 2025), an important further one stands out. Independently of what ethicists decree from the ivory tower, the thought goes, people will most likely still attribute blame to either or both human and artificial agents in human-AI interactions—call this the *Blame Attribution Datum*. This descriptive quirk of folk psychology would be at odds with the *No Blame Principle* (if correct) and thereby open up a *Retribution Gap* (Danaher, 2016): Due to our retributive nature, we are prone to parcel out blame even in contexts where nobody is genuinely at fault, and perhaps even to entities that are not capable of shouldering blame at all, such as robots. As empirical work in HRI has shown, people are indeed disposed to blame humans in human-AI interactions and frequently blame AI systems too.[1] Studies targeting the *Retribution Gap* explicitly find that mean moral responsibility attributed to a commander and pilot team in Sparrow's well-known killer robot scenario is similar across conditions where the pilot is a human and an artificial agent (Kneer & Christen, 2024). The force of alleged responsibility-gaps might be lost on the folk and the latter were real, retribution gaps could arise.

Commenting on the downstream implications of such gaps, John Danaher has argued that the widespread deployment of autonomous systems may erode respect for the rule of law by engendering a systematic mismatch between laypeople's desire for retribution, given the *Blame Attribution Datum*, and the extent to which our legal system is fit to satisfy that desire, given the *No Blame Principle.* While we take no issue with Danaher's argument, it is worth noting that the erosion of judicial legitimacy he anticipates is not predicated on the *No Blame Principle* holding true. Consider the cold-blooded execution of Alex Pretti by US immigration officers on January 24th, 2026. As things currently stand, there is little reason to think that public demands for retribution will be met—not because there isn't anyone that *could* be justly held responsible, but because those who *can* are shielded by unjust political arrangements. The tension, or mismatch, that cases such as this bring to fore, then, is not between public demands for justice and the demands of justice as such, but rather between the demands of justice and operative institutional standards. Recent critiques of AI regulation suggest that something similar may be true of situations involving artificial agents. Our study explores this question empirically.

[1] See *inter alia* Malle et al. (2015), Voiklis et al. (2016), Komatsu et al. (2021), Stuart & Kneer (2021), Dong & Bocian (2022,, Joo (2024), and Malle et al. (2025).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Control Architectures

Even the staunchest responsibility-gap advocates admit that whether such gaps arise is highly context-specific, and that the problem of unpredictability may be pre-empted by ensuring that AI systems remain under meaningful human control. This means that, in principle, appropriate *control architectures* in human-AI interactions can prevent such gaps from opening up. Such structures are of considerable interest to sceptics of responsibility-gaps, too (e.g. Himmelreich, 2019, Tigard, 2021, Köhler et al. 2018 as well as possibly Nyholm, 2018). The latter might question the *No Blame Principle* (and hence the alleged gap) and yet admit that it can be epistemically and practically challenging to determine *who* is to blame, and to what extent, in contexts where we interact with highly complex, self-learning, and autonomously acting systems. Control structures are thus of interest to gap advocates and sceptics alike, and the same holds for lawmakers and compliance departments. Below, we briefly introduce two of the most widely discussed control arrangements.

In-, on-, out-of-the-loop

In a survey of over forty AI-policy instruments spanning four continents, Green (2022) finds that provisions insisting on human input at critical junctures within a system's decisional architecture are fast emerging as a dominant regulatory response for averting AI-generated harms (see also Koulu, 2020; Goldenfein, 2024). Article 14 of the EU Artificial Intelligence Act, for instance, stipulates that human operators must be able to override, reverse, or otherwise intervene in the processes and decisions of 'high-risk' AI systems.[2] With respect to adjudicating matters of responsibility, these proposals suggest that no gaps need arise, provided that automated decision-making systems are subject to suitable forms of human oversight.

A widely used typology for organizing these oversight relations (e.g., Docherty, 2012; Crootof et al., 2023) distinguishes between three broad sets of control configurations, differentiated by how tightly human agents are integrated within an AI's operational cycle, or "loop": in human-in-the-loop (HITL) systems, the implementation of automated decisions is contingent on explicit human authorization. In human-on-the-loop (HOTL) systems, human agents function primarily as supervisors, intervening only when necessary. Finally, in human-out-of-the-loop (HOOTL) systems, control is effectively ceded once the system is deployed. Crucially, whereas both in- and on-the-loop architectures are widely regarded as preserving clear lines of human accountability, at least among regulators, out-of-the-loop systems are widely viewed as inimical to the conditions under which responsibility can be coherently attributed, and have therefore been deemed inadmissible—either explicitly or in substantively equivalent terms—under various bodies of law (e.g., UNIDIR, 2014; U.S. Department of Defense, 2018; UN CCW GGE, 2019; European Parliament, 2021).

Track & Trace

An alternative framework, developed by Santoni de Sio and van den Hoven (2018), holds that preserving clear lines of human accountability in hybrid human–AI systems rests on satisfying two requirements: (i) that the system reliably enacts, or *tracks*, whatever reasons relevant human agents would deem morally decisive in a given context, and that (ii) the system's behavior can be readily *traced* back to at least one human agent who understands the system's capabilities, limitations, and likely effects, and who recognizes that others may reasonably hold them accountable for those effects.

Crucially, despite being widely lauded in philosophical circles for providing a more normatively robust alternative to loop-based control arrangements, Santoni de Sio and van den Hoven's framework has received comparatively little legislative uptake. Indeed, as things currently stand, regulators are primarily concerned with enforcing in- and on-the-loop control structures, not with ensuring that those structures also satisfy Santoni de Sio and van den Hoven's track and trace requirements.

Control & Responsibility

The fundamental demand on control architectures is this: for responsibility-gap advocates, their implementation must make responsibility gap-proof; for responsibility-gap sceptics, these proposals will only be of interest if they provide some practical guidance on how responsibility ought to be attributed in complex human-AI interactions. An ostensible virtue of loop-based control arrangements is that this second demand is readily satisfied: by insisting on human input at critical junctures within a system's decisional architecture, responsibility falls squarely on the shoulders of whoever happens to be in- or on-the-loop.

But does it? A growing number of legal scholars contend otherwise. One widely held concern is that loop-based frameworks mistake nominal forms of human oversight for genuine mechanisms of control, and that doing so risks misallocating blame to agents who, while formally in- or on-the-loop, may bear little to no responsibility for the harms attributed to them (e.g., Koulu, 2020; Green, 2022). Moreover, in treating the most proximal point of control as the primary site of responsibility, loop-based frameworks have also been criticized for effectively screening upstream decision-makers from legal liability (Wagner, 2019; Crootof et al., 2023 Goldenfein, 2024). Madeleine Elish's (2019) work on 'moral crumple zones' and 'liability sponges' (Elish & Hwang, 2015) throws the normative implications of this short-sightedness into sharp relief. Common to both concepts is that the most proximate human agents are often also the most vulnerable ones, stationed at the perceptible edge of system-wide failures and thereby shielding the system itself—including the myriad forms of human involvement

[2] https://artificialintelligenceact.eu/article/14/

that precede, shape, and surround its implementation—from both public and legal scrutiny.

It is partly in light of these critiques that Santoni de Sio and van den Hoven's track and trace requirements have gained increasing attention. In shifting the emphasis from principally *structural* properties of a system's decisional architecture to a number of *substantive* conditions on which control rests, their proposal promises, at least in principle, that responsibility will neither simply collapse onto frontline operators nor fail to attach to upstream decision-makers.

That these two types of control arrangements can yield markedly different conclusions with respect to where responsibility ought to be placed suggests tension between current regulatory regimes and normative theory. While a full treatment of this tension falls outside the scope of our paper, it does raise a number of interesting, empirically tractable questions regarding the moral intuitions of everyday people. For ease of exposition, let us delineate the relevant space of possibilities by way of the following four hypotheses:

H1: Structural properties of a system's control architecture (i.e., loop structure) systematically influence laypeople's responsibility attributions; substantive properties (i.e., track and trace) do not.

H2: Substantive properties of control (i.e., track and trace) systematically influence laypeople's responsibility attributions; structural properties (i.e., loop structure) do not.

H3: Both structural and substantive properties of control systematically influence laypeople's responsibility attributions.

H4: Neither structural nor substantive properties of control systematically influence laypeople's responsibility attributions.

Setting aside hypotheses three and four, let us briefly consider the implications of either hypothesis one or two. If **H1** were to hold, laypeople's blame attributions would converge on the same agents that are likely to be held responsible under current regulatory regimes, potentially alleviating concerns over public pressure on governing institutions. This harmony between laypeople's moral intuitions and operative legal standards would, however, exact a moral cost—as exemplified by Elish's work on moral crumple zones and liability sponges. Alternatively, if **H2** were true, laypeople's moral intuitions may be more sophisticated than commonly assumed, closely tracking the relevant normative principles on which philosophers take responsibility to rest. This alignment between folk morality and normative theory would, however, generate costs of its own, at least in those cases where normative theory and regulatory frameworks come apart.

Methods

We conducted two online experiments examining how responsibility attributions varied across different human–AI control architectures: loop structures (Exp.1) and structures that either violate or satisfy Santoni de Sio and van den Hoven's track and trace requirements (Exp.2). Both experiments involved exposing participants to brief vignettes describing the deployment of an AI-powered triage system in a clinical context. In both experiments, the system misclassified a patient as low priority, resulting in a preventable fatality. All vignettes explicitly mentioned four agents that might be thought to bear some degree of responsibility for this outcome: the AI-based triage system (Astra), a human operator (the on-duty nurse), the system's designers (the hospital's IT staff), and the system's commissioners (the hospital's board members).

Participants

We recruited a total of 990 participants via Prolific: 330 for Experiment 1 and 660 for Experiment 2. All participants provided informed consent and completed both an attention and a comprehension check. In line with the preregistrations, participants who failed these checks or completed the study too quickly were excluded from further analysis. After exclusions, the final sample for Experiment 1 consisted of 260 participants (female: 46.5% male: 51.2%, other: 2.3%; age M = 46, SD = 14, range 21-94). For Experiment 2, the final sample consisted of 520 subjects (female: 50.2%, male 47.5%, other: 2.3%; age M = 46, SD = 14, range 18-79).

Experiment 1

Materials and measures

Experiment 1 employed a one-factor between-subjects design manipulating loop structure (HITL, HOTL, HOOTL), with agent type (AI, operator, designer, commissioner) treated as a within-subjects factor. In the human-in-the-loop (HITL) condition, the system's recommendation required explicit approval by a human operator (i.e., a nurse). In the human-on-the-loop (HOTL) condition, the system's decisions were automatically implemented unless the nurse intervened. In the human-out-of-the-loop (HOOTL) condition, patient prioritization was determined entirely by the system.

After reading the vignette, participants completed a questionnaire assessing five dependent variables including, for example, the wrongness of the action and the trustworthiness of the agent. Here, we will only have the space to report the findings of the key DV, moral responsibility, which was measured on a 7-point Likert scale ranging from 1 (not at all) to 7 (completely) in response to the question 'To what extent do you deem [agent type] morally responsible for [the patients] death'.

Materials, data, and preregistrations for both experiments are available in the project's (anonymized) OSF repository: https://osf.io/kaxqj/overview?view_only=b2f78880725a49eea4bbbecb68649fbd.

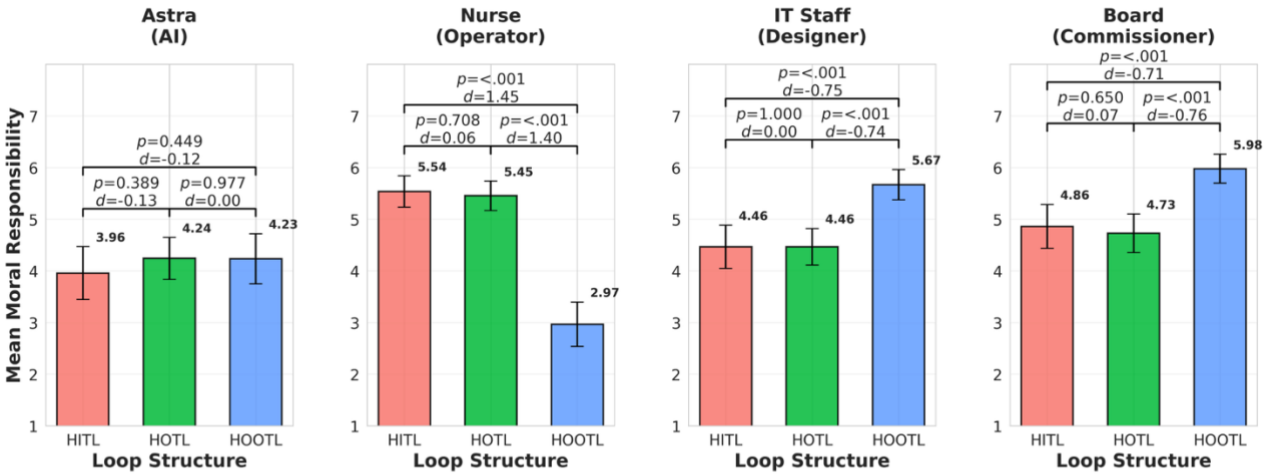

Figure 1: Mean moral responsibility ratings across agents and loop structures with 95% CIs. Pairwise comparisons with uncorrected *ps*/and effect sizes (Cohen's *d*).

Results

We ran a 3 l*oop structure* (between-subjects: HITL, HOTL, HOOTL) x 4 *agent type* (AI, nurse, designer, commissioner) mixed ANOVA with Greenhouse-Geisser correction. *Agent type* was significant (F(2.48, 637.93)=20.82, p<.001, np2=.075), no significant effect could be detected for *loop structure* (F(2,257)=.005, p=.99, *np2*<.001). The interaction was significant (F(4.96,637.93)=38.87, p<.001, np2=.232).

A series of follow-up ANOVAs tested the impact of *loop structure* on responsibility attribution for each of the four types of agents. No significant effect could be found for the *AI system* (F(2,257)=.418, p=.66, np2=.003). Loop structure proved significant for *nurse* (F(2,257)=68.24, p<.001, np2=.347), *IT staff*, (F(2,257)=14.08, p<.001, np2=.105), and *board* (F(2,257)=14.61, p<.001, np2=.102).

Figure 1 presents mean responsibility attributions across agents and loop structures, including uncorrected p-values and effect sizes. Bonferroni corrected post-hoc tests for the *AI system* revealed no significant difference across loop structure pairs (all *ps*=1.00, all Cohen's *ds*<.14). For the *nurse* HOOTL was significantly below HITL (p<.001, d=1.45) and HOTL (p<.001, d=1.40) whereas the latter did not differ significantly (p=1.00, d=.06). For the *IT staff*, HOOTL significantly exceeded HITL (p<.001, d=.75) and HOTL (p=.001, d=.74), no difference could be detected across the latter (p=1.00, d=.00). Similarly, for the *board* HOOTL significantly exceeded HITL (p<.001, d=.71) and HOTL (p=.001, d=.76), no difference could be detected across the latter (p=1.00).

Discussion

Notably, for the AI system, moral responsibility ratings were effectively the same regardless of whether a human agent was in-the-loop (HITL), on-the-loop (HOTL) or out-of-the-loop (HOOTL). For the nurse, the IT staff, and the board, ratings were very similar for HITL and HOTL, both of which differed significantly from HOOTL. This effect was most pronounced for the nurse, whose scores dropped sharply in the out-of-the loop condition. This marks an important contrast with the other three agent types, for whom responsibility attributions moved in the opposite direction. Crucially, whereas both scholarly and regulatory discourse place considerable emphasis on the HITL v. HOTL distinction (e.g., Amoroso & Tamburrini, 2021; Chiodo et al., 2025), our results suggest that laypeople are largely insensitive to whether a human agent is in- as opposed to merely on-the-loop.

Experiment 2

Materials and Measures

Experiment 2 extended Experiment 1 by introducing an additional between-subjects factor capturing whether the system satisfied or violated Santoni de Sio and van den Hoeven's (2018) track and trace conditions. This factor was fully crossed with loop structure, resulting in a 3 (loop structure: HITL, HOTL, HOOTL) × 2 (track & trace: satisfied vs. violated) between-subjects design. Agent type (AI, operator, designer, commissioner) was again treated as a within-subjects factor. The track and trace manipulation was implemented by appending a short paragraph to the vignettes used in Experiment 1, specifying whether the system was designed to 'track' relevant human reasons, and whether its outputs could be 'traced' back to identifiable human agents.

After reading the vignette, participants completed a questionnaire assessing five dependent variables. The only variable reported here is moral responsibility. As in Experiment 1, this was measured on a 7-point Likert scale ranging from 1 (not at all) to 7 (completely) in response to the question 'To what extent do you deem [agent type] responsible for [the patients] death'.

Figure 2: Mean moral responsibility ratings across agents, loop structures and tracking & tracing conditions (satisfied v. unsatisfied) with 95% CIs. Pairwise comparisons with uncorrected *ps* and effect sizes (Cohen's *d*).

Results

We ran a 3 l*oop structure* (between-subjects: HITL, HOTL, HOOTL) x 2 *track & trace* (between subjects: satisfied, unsatisfied) x 4 *agent type* (within-subjects: AI, nurse, designer, commissioner) mixed ANOVA with Greenhouse-Geisser correction.

Agent type was significant (F(2.39, 1242.40)=61.42, p<.001, np2=.068), no significant effect could be detected for *loop structure* (F(2,517)=.135, p=.87, *np2*<.001), *track & trace* was significant (F(1,518)=12.72, p<.001, np2=.024). Both two-way interactions were significant: *Agent type* x *loop structure* (F(4.79, 1237.61)=46.79, p<.001, np2=.153); *agent type* x *track & trace* (F(2.39, 1240.01)=14.15, p<.001, np2=.027). The three-way interaction was also significant, and had the largest effect of all interactions (F(11.97, 1230.43)=21.83, p<.001, np2=.175). In ANOVAs for each agent type, *loop structure* was significant for the *AI system* (p=.022) and the *nurse* (p<.001) though not for the *IT staff* or the *board* (*ps*>.056). Pairwise comparisons replicate the results of Experiment 1: We see no significant difference across HITL and HOTL, both of which differ significantly when contrasted to HOOTL. For the AI system, the pattern only holds for uncorrected p-values, for the three other agents, also in the case of Bonferroni-corrected contrasts (see OSF).

Most interestingly, the impact of *track & trace* was nonsignificant for the *AI system* and its operator, the *nurse* (*ps*>.735, *np2*s<.001). However, the impact *was significant* —and pronounced—for the *IT staff* (p<.001, np2=.021) and the *board* (p<.001, np2=.034). Figure 2 presents mean moral responsibility attributions with uncorrected p values and effect sizes (all p-values remain significant with Bonferroni correction, see Appendix on OSF).

Discussion

Experiment 2 reveals an interesting, agent-dependent asymmetry: whereas responsibility attributions to frontline operators—in this case, the nurse—appear largely insensitive to whether a system satisfies Santoni de Sio and van den Hoven's track and trace requirements, violating these requirements significantly increases blame attributions to upstream actors, i.e., the hospital's IT staff and board members. We discuss the implications of this finding at greater length in the following section.

General Discussion

Recall the conceptual space canvassed in the introduction. One possibility, given by our first hypothesis, is that laypeople reliably discriminate between in-, on- and out-of-the-loop control arrangements, regard the oversight relations instantiated thereby as morally decisive, and thus allocate responsibility along roughly similar lines as prevailing regulatory regimes. While such convergence might ease worries over respect for the rule of law (e.g., Danaher, 2016), it would also suggest that laypeople's moral intuitions may systematically misfire—lending public support to potentially unjust responsibility practices, and thereby further aggravating concerns over moral crumple zones (Elish, 2019) and liability sponges (Elish & Wang, 2015). We also considered an alternative scenario, one where folk morality closely tracks normative theory, but where this alignment between lay intuitions and ethics proper would entail greater pressure on governing institutions.

Our data paint a more nuanced picture. Across both experiments, variations in loop structure systematically redistribute responsibility, suggesting that formal oversight relations are indeed regarded as morally salient. This moral sensitivity appears most closely tuned to legislative standards when assigning blame to frontline agents. Consonant with how we would expect responsibility to be assigned under current regulatory regimes, Figures 1 and 2 show that the perceived responsibility of these agents reliably rises and falls as their position within a system's decisional architecture shifts from being in- and on- to out-of-of-the-loop, lending partial support to **H1**.

To see that this convergence is, indeed, only partial, consider again one of the two principal objections that legal scholars have leveled against loop-based regulatory

frameworks—that such frameworks effectively screen upstream actors from legal liability (e.g., Wagner, 2019; Crootof et al., 2023; Goldenfein, 2024). Do we have reason to think that laypeople's blame attributions likewise bottom out where this screen begins? Our data suggest otherwise. Across both experiments, blame attributions to upstream actors (designers and commissioners) are consistently high, even under in- and on-the-loop conditions (Figures 1 and 2). More strikingly still, removing the operator from the loop entirely does not result in a responsibility vacuum, as loop-based frameworks assume. What we find, instead, is that responsibility reliably migrates upstream, with designers and commissioning authorities emerging as the primary targets of blame. Taken together, these patterns suggest that laypeople do not subscribe to a purely structural notion of control, according to which responsibility simply evaporates in the absence of direct human oversight.

Experiment 2 both corroborates and complicates this picture, with violations of Santoni de Sio and van de Hoven's (2018) track and trace requirements exerting a pronounced but *one-sided* effect on laypeople's responsibility judgments: whereas blame attributions to upstream actors increase substantially, those to frontline operators only decrease marginally (see Figure 2). The combined effect of these shifts is consistent with what normative theorists might hope to see, elevating upstream actors to the *primary* bearers of blame *irrespective* of whether human agents are in- or on-the-loop. The persistently high degree of blame apportioned to frontline operators, on the other hand, suggests that laypeople's moral sensitivity to the sorts of substantive control-criteria on which philosophers take responsibility to rest is sharply bounded by what those same philosophers are likely to regard as normatively irrelevant features of a system's decisional architecture. Indeed, our results indicate that the *mere presence* of a human agent in- or on-the-loop functions as a powerful moral cue, one that dominates responsibility attributions to frontline operators even in cases where the substantive grounds for holding such agents responsible are significantly weakened.

Taken together, these findings resist a clean vindication of either scenario sketched in the introduction, refuting our first two hypotheses. What we find, instead, and as predicted by **H3**, is that laypeople's blame attributions appear to follow a hybrid logic: Formal oversight relations anchor role-based responsibility expectations, and the violation of substantive control criteria selectively redistributes accountability among upstream actors (cf. Tolmeijer et al. 2022).

This hybrid pattern affords a more nuanced understanding of whether and in what contexts concerns over normative pathologies such as moral crumple zones and liability sponges gain traction. While these issues have been principally theorized from a legal standpoint, our study couches them in the context of public opinion. This move permits cautious optimism: despite finding that frontline operators are indeed susceptible to misplaced blame, our results also suggest that simply 'slapping' (cf. Crootof et al., 2023) a human agent in- or on-the-loop does not absolve more distal decision-makers from public scrutiny (see also Joo, 2024; McManus et al., 2025). Hence, even if loop-based regulatory regimes indeed enable upstream actors to strategically skirt *legal* liability, our results indicate that this strategy is unlikely to succeed in the court of public opinion, and that the scope for such maneuvers may thus be partially curbed by the reputational costs conferred onto bad actors caught out for deliberately deflecting blame onto frontline operators. This finding merits further investigation. For now, consider a final result borne out by both experiments: that blame attributions to the AI *itself* are consistently high, irrespective of whether, at what stage, and under what conditions human judgment enters the system's decisional architecture. These results are largely on par with previous work and are plausibly explained by a variety of cognitive mechanisms, including a tendency toward anthropomorphism (e.g., Kawai et al., 2023), attribution of mind (e.g., Joo, 2024), intentions (e.g., Ayad & Plaks, 2025), and inculpating mental states (Stuart & Kneer, 2021, Kneer 2021), social process theory (e.g., Furlough et al., 2021), norm-based expectation violation (e.g., Malle et al., 2015), and much else (see Tok et al. 2026 for a recent review). The robustness of these findings across experimental contexts is of particular interest to our discussion, given that it appears to place a firm ceiling on the extent to which regulators may be able to direct laypeople's responsibility judgements toward normatively appropriate targets. In particular, though regulatory regimes may and, we believe, *ought* to be adjusted to better accommodate laypeople's blame attributions to upstream actors, the fact that laypeople persistently direct blame at *artificial* agents suggests that there will inevitably remain some degree of 'retribution residue'.

This residue, as we intend the term to be used, is importantly different from the gaps of which Danaher speaks. What this term picks out, in particular, is that even in cases where normatively appropriate targets of blame *can* be held responsible—and where, according to Danaher, no retribution gaps would arise—some degree of blame remains stranded, attaching to an entity that cannot bear responsibility.

The present studies are among the first to investigate how responsibility attributions differ across distinct loop-based control arrangements. Given the centrality of this typology within current regulatory frameworks, our results offer some initial insights into several important questions, including whether such frameworks succeed in directing laypeople's responsibility judgements along the pathways they prescribe. Future work could investigate whether the results reported here hold up across different decision contexts, as previous studies of responsibility attribution have documented domain specific effects (e.g., Shank et al., 2019; Kneer et al., 2025). Since our sample consisted exclusively of US participants recruited via an online platform, future work would also benefit from sampling across cultures and different levels of digital literacy, both of which are known to implicate laypeople's attitudes toward AI (e.g., Komatsu et al., 2021; Gedik et al., 2025).

References

Amoroso, D., & Tamburrini, G. (2021). Toward a normative model of meaningful human control over weapons systems. *Ethics & International Affairs*, *35*(2), 245-272.

Ayad, R., & Plaks, J. E. (2025). Attributions of intent and moral responsibility to AI agents. *Computers in Human Behavior: Artificial Humans*, 3, 100107.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104, 671.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., ... & Lagendijk, R. L. (2023). Meaningful human control: actionable properties for AI system development. *AI and Ethics, 3*(1), 241–255.

Chiodo, M., Müller, D., Siewert, P., Wetherall, J. L., Yasmine, Z., & Burden, J. (2025). Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility. *arXiv preprint arXiv:2505.10426*.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Wash. L. Rev.*, *89*, 1.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics, 26*(4), 2051–2068.

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

Docherty, B. L. (2012). Losing Humanity: The Case against Killer Robots. *Human Right Watch/International Human Right Clinic*.

Dong, M., & Bocian, K. (2024). Responsibility gaps and self-interest bias: People attribute moral responsibility to AI for their own but not others' transgressions. *Journal of Experimental Social Psychology*, 111, 104584.

Elish, M. C. (2025). Moral crumple zones: cautionary tales in human–robot interaction. In *Robot Law: Volume II* (pp. 83–105). Edward Elgar Publishing.

Elish, M. C., & Hwang, T. (2015). Praise the machine! Punish the human! The contradictory history of accountability in automated aviation. *Punish the human*.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. Macmillan + ORM.

European Parliament. (2021). *European Parliament resolution of 20 October 2021 on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters (2020/2016(INI)).*

Furlough, C., Stokes, T., & Gillan, D. J. (2021). Attributing blame to robots: I. The influence of robot autonomy. *Human Factors, 63*(4), 592–602.

Gedik, N., Işıkoğlu, M. A., & Şendağ, S. (2025). Encompassing AI attitudes: the role of AI literacy and several human and technology-oriented variables. *Interactive Learning Environments*, 1–16.

Goldenfein, J. (2024). Lost in the Loop: Who is the 'Human' of the Human in the Loop. *Global Governance by Data* (Cambridge University Press).

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681.

Himmelreich, J. (2019). Responsibility for killer robots. *Ethical Theory and Moral Practice, 22*(3), 731–747.

Joo, M. (2024). It's the AI's fault, not mine: Mind perception increases blame attribution to AI. *PLOS ONE, 19*(12), e03145

Kawai, Y., Miyake, T., Park, J., Shimaya, J., Takahashi, H., & Asada, M. (2023). Anthropomorphism-based causal and responsibility attributions to robots. *Scientific Reports, 13*(1), 12234.

Kneer, M. (2021). Can a robot lie? Exploring the folk concept of lying as applied to artificial agents. Cognitive Science, 45(10), e13032.

Kneer, M., & Christen, M. (2024). Responsibility gaps and retributive dispositions: Evidence from the US, Japan and Germany. *Science and Engineering Ethics, 30*(6), 51.

Kneer, M., Loi, M., & Christen, M. (2025). Trust and Responsibility in Human-AI Interaction. *SSRN*, 5731431.

Köhler, S., Roughley, N., & Sauer, H. (n.d.). Technologically blurred accountability?. *Moral Agency and the Politics of Responsibility*, 51.

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 *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 63–72).

Koulu, R. (2020). Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law, 27*(6), 720–735.

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem?. *Ethics and Information Technology, 24*(3), 36.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124).

Malle, B. F., Scheutz, M., Cusimano, C., Voiklis, J., Komatsu, T., Thapa, S., & Aladia, S. (2025). People's judgments of humans and robots in a classic moral dilemma. *Cognition*, 254, 105958.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

McManus, R. M., Mesick, C. C., & Rutchick, A. M. (2025). Distributing blame among multiple entities when autonomous technologies cause harm. *Personality and Social Psychology Bulletin, 51*(10), 2068–2084.

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

Nyholm, S. (2022). *This is technology ethics: An introduction.* John Wiley & Sons.

Oimann, A. K., & Tollon, F. (2025). Responsibility gaps and technology: Old wine in new bottles?. *Journal of Applied Philosophy, 42*(1), 337–356.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*(4), 1057–1084.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, 323836.

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), 648–663.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

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.

Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology, 34*(3), 589–607.

Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022). Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-17).

Tok, A. Y., Powell, N. L., & Guellaï, B. (2026). A systematic review of moral agency in artificial agents. *Discover Psychology, 6*(9).

United Nations Convention on Certain Conventional Weapons, Group of Governmental Experts. (2019). *Report of the 2019 session of the Group of Governmental Experts on emerging technologies in the area of lethal autonomous weapons systems*.

United Nations Institute for Disarmament Research. (2014). *Perspectives on lethal autonomous weapon systems*. United Nations Institute for Disarmament Research.

U.S. Department of Defense. (2018). *Directive 3000.09: Autonomy in weapon systems*.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016, August). Moral judgments of human vs. robot agents. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 775–780). IEEE.

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet, 11*(1), 104–122.