In the loop, out of sync: moral cognition in human-robot interactions

Anonymized for review

The ubiquitous integration of AI-powered systems in morally consequential decision-making procedures has been shadowed by an expanding catalogue of harms [e.g., 1,2,3]. 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 [4]. Who, for instance, should be held accountable when an AI-powered triage tool mistakenly classifies an incoming patient as low priority, resulting in a preventable fatality? According to recent responses, these sorts of cases may be less intractable than they initially appear [e.g., 5,6,7]. Köhler and colleagues [8; see also 9], 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 renders it in principle impossible to *justly* attribute blame, lest the actions of autonomous systems are *suitably constrained* [10,11].

Indeed, even the staunchest responsibility-gap advocates admit that whether such gaps arise is highly context-specific, and that the problem of unpredictability may be preempted by ensuring that automated systems remain under meaningful human control. This means that, in principle, appropriate *control architectures* in human-AI interactions can prevent responsibility-gaps from opening up. Such structures are of interest to sceptics of responsibility-gaps, too. The latter [e.g., 5,6,8] might question whether responsibility-gaps are real 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 systems. Control structures are thus of interest to gap advocates and sceptics alike, and the same holds for lawmakers and compliance departments.

We present two experiments (n=780) investigating how laypeople attribute blame under two of the most widely discussed control architectures: loop-based control structures [12,13], and control structures that either violate or satisfy Santoni de Sio's and Van Den Hoven's [14] 'track and trace' requirements. A widely used typology for organizing loop-based oversight relations distinguishes between three broad types of control configurations: human-in-the-loop (HITL) systems, where automated decisions are contingent on human authorization; human-on-the-loop (HOTL) systems, where human agents intervene only when necessary; and human-out-of-the-loop (HOOTL) systems, where control is effectively ceded once the system is deployed. Crucially, where loop-based approaches index control to principally *structural* properties of human oversight, Santoni de Sio and Van Den Hoven's 'track and trace' requirements specify several *substantive* conditions on which control rests: (i) that the system reliably *tracks* whatever reasons relevant human agents would deem morally decisive in a given context, and that (ii) the system's behavior can be *traced* back to at least one human agent who understands the system's capabilities, limitations, and likely effects.

Whether either of these approaches fully succeeds in the difficult task of *justly* assigning blame in human-AI interactions is an important and widely debated question [12,13]. It is, however, not the only question we should ask. Effective AI governance does not hinge solely on the normative adequacy of the proposed control structures; it also depends on whether these structures resonate with lay intuitions about who is, and ought to be, responsible. Governance regimes that assign responsibility in ways that strike ordinary people as unintuitive or opaque risk being perceived as arbitrary or unfair, thereby undermining compliance and legitimacy [15,16].

Experiments

Participants & Methods: We conducted two online experiments examining how responsibility attributions varied across different human–AI control architectures: *loop structures* (between-subjects: HITL, HOTL, HOOTL, *n*=260) and structures that either violate or satisfy Santoni de Sio and Van Den Hoven's *track and trace* requirements (between-subjects: 3 (HITL, HOTL, HOOTL) x 2 (track & trace: satisfied v. unsatisfied, *n*=520)). Across both experiments, participants saw vignettes describing the deployment of an AI-powered triage system in a medical context that misclassified a patient, resulting in a preventable fatality. All vignettes 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 (nurse), the system's designers (the hospital's IT staff), and the system's commissioners (the hospital's board).

Results: In Exp.1, moral responsibility ratings were effectively the same across in-the-loop (HITL), on-the-loop (HOTL), and out-of-the-loop (HOOTL) for the AI system, see Fig.1. Ratings were very similar for HITL and HOTL for the nurse, the IT staff, and the board. For these three actors, both HITL and HOTL differed significantly from HOOTL (all *ps*<.001), with very large effect sizes for the nurse (Cohen's *ds*>1.39), and medium-large effect sizes for IT staff and board (*ds*>.73). Interestingly—and importantly—whereas both scholarly and regulatory discourse place considerable emphasis on the HITL v. HOTL distinction [12, 13], laypeople appear largely insensitive to whether a human agent is in- as opposed to merely on-the-loop.

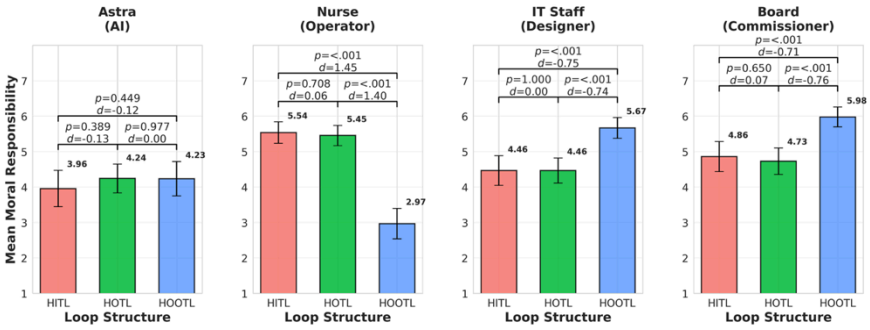

Figure 1: Mean moral responsibility ratings across agents and loop structures with 95% CIs. Pairwise comparisons with uncorrected *ps* and effect sizes (Cohen's *d*).

Beyond replicating the general patterns observed in Exp.1, Exp.2 revealed an interesting, agent-dependent asymmetry (see Fig.2): 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 IT staff (*p*<.001, np2=.021) and the hospital's board (*p*<.001, np2=.034). This finding suggests that although design failures systematically shift blame higher up the decision-making chain, they do not mitigate the perceived importance of operational proximity.

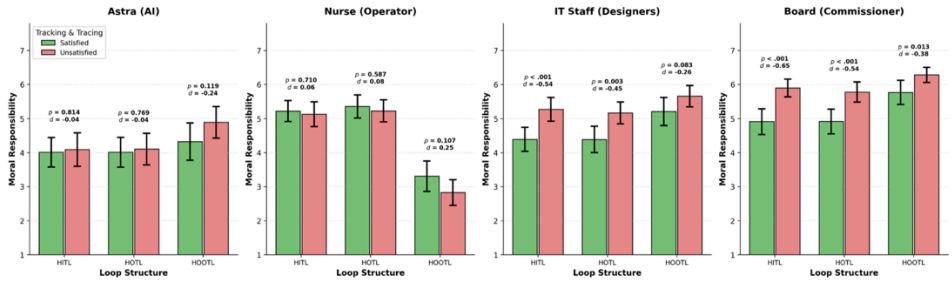

Figure 2: Mean moral responsibility ratings across agents, loop structures and tracking & tracing conditions (satisfied v. unsatisfied) with 95% CIs. Pairwise comparisons with uncorrected *ps* and effect sizes (Cohen's *d*).

In the talk, we will present additional scenarios and control architectures (total n=3700) and draw implications for governance and legal regulation. We will also introduce a framework that treats meaningful human control not only as a normative ideal, but as an empirically informed set of control principles that accommodate folk psychology.

References

[1] Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, 89, 1–33.

[2] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671–732.

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

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

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

[6] Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *S.&E. Ethics, 26*(4), 2051–2068.

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

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

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

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

[11] Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 2*4(1), 62–77.

[12] Jones, M. L. (2015). The ironies of automation law: Tying policy knots with fair automation practices principles. *Vand. J. Ent. & Tech. L.*, *18*, 77.

[13] Crootof, R., Kaminski, M. E., Price, W., & Nicholson, I. I. (2023). Humans in the Loop. *Vand. L. Rev.*, *76*, 429.

[14] 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.

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

[16] Tyler, T. R. (2006). Psychological perspectives on legitimacy and legitimation. *Annu. Rev. Psychol.*, *57*(1), 375-400.