Public Perspectives on Robot Intervention: Insights from a Low-Supervision Decision Exhibit

Arjita Mital*
Karlsruhe Institute of Technology
Karlsruhe Germany

Felix Gnisa
Karlsruhe Institute of Technology
Karlsruhe Germany

Utku Norman
Karlsruhe Institute of Technology
Karlsruhe Germany

Nora Weinberger
Karlsruhe Institute of Technology
Karlsruhe Germany

Abstract

This Late Breaking Work presents a low-threshold, unsupervised public exhibit designed to explore how non-expert audiences imagine and negotiate future human–robot interactions in ethically charged everyday situations. The exhibit, installed in Karlsruhe, Germany, invited participants to engage with four dilemma-based scenarios where participants were prompted to decide how a social robot should act confronting questions of moral delegation and machine agency. The activity generated rich, situated reflections on responsibility, safety, care, and the limits of automation. Findings reveal context-dependent expectations that balance efficiency against dignity, human judgment, and relational preservation, shaped by perceived stakes, social context, and the specific embodiment of the robot involved. Through this we demonstrate how minimally supervised participatory formats can surface normative expectations and support inclusive, responsible robot design.

CCS Concepts

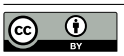
• **Human-centered computing** → **Empirical studies in HCI**; *Interaction design theory, concepts and paradigms*; • **Social and professional topics** → *Socio-technical systems.*

Keywords

Public engagement, robotic decision-making, moral delegation, participatory futures, ethical dilemmas

ACM Reference Format:
Arjita Mital, Felix Gnisa, Utku Norman, and Nora Weinberger. 2026. Public Perspectives on Robot Intervention: Insights from a Low-Supervision Decision Exhibit. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3776734.3794569

*This work was funded by Baden-Württemberg Ministry of Science, Research and Art (MWK), using funds from the state digitalisation strategy digital@bw. Institute for Technology Assessment and Systems Analysis (ITAS). Corresponding author: arjita.mital2@kit.edu



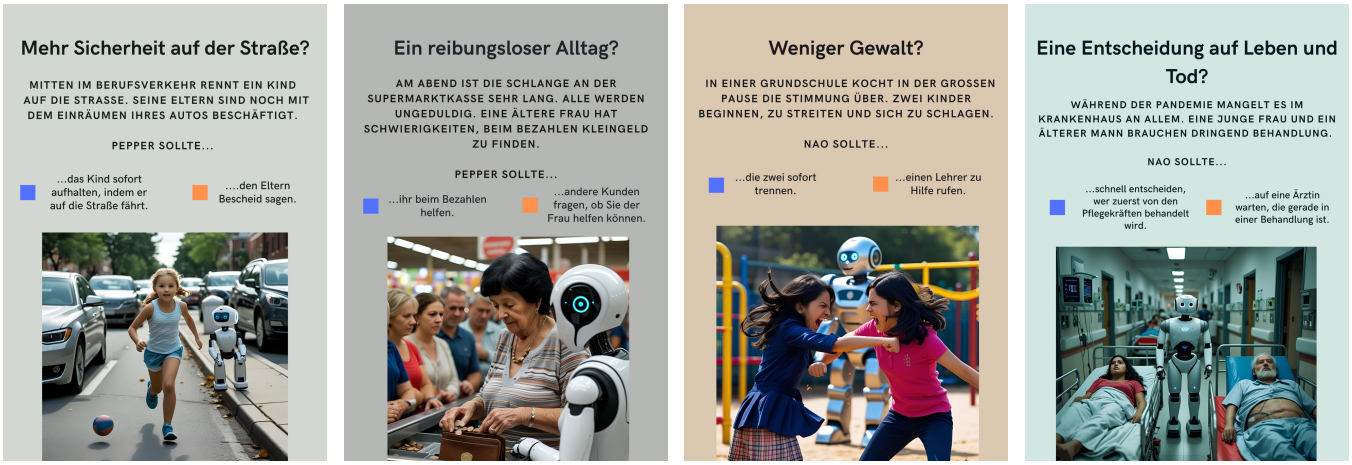
HRI Companion '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3794569

1 Introduction and Motivation

The rapid advancement of robotics and autonomous AI systems raises urgent questions about their role in critical decision-making. Devolving responsibility to robots presents significant risks, as this shift often entails a loss of human judgment and accountability. Researchers have warned of the "control problem" [11], where humans become overly reliant on machine decisions, potentially compromising safety, while others highlight that autonomous robots' decisions may not be predictable or fully explainable, reducing trustworthiness [3]. As these technologies expand into domains affecting human safety, health, and lives, experts recommend maintaining meaningful human oversight: preserving human expertise within decision loops [6], designing inherently interpretable models [8], and implementing human-in-the-loop systems to ensure accountability [7]. These approaches underscore the imperative of active human involvement in shaping robotic futures. While mutual shaping approaches have demonstrated the value of involving publics in imagining robotic futures [9, 10], there is limited knowledge about how laypeople engage with robotic decision-making itself, particularly in morally ambiguous or socially sensitive situations. Contemporary approaches to robotic decision-making, ranging from reinforcement learning [2] to probabilistic models such as Bayesian networks and Markov decision processes [12], as well as deep learning architectures [4], primarily reflect technical optimisation goals. At present, opportunities for non-experts to shape or interrogate these decision processes remain limited, and questions of public legitimacy, perceived appropriateness, or moral acceptability of robotic actions are often left unexplored. This disconnect leaves a growing gap between the design of autonomous behaviours and public values, expectations, and concerns.

This Late Breaking Work presents an initial attempt to address this gap through a low-threshold, unsupervised public exhibit designed to engage diverse publics in reasoning about robotic intervention. Rather than asking abstract questions about robotics in society, the format placed participants in concrete dilemma scenarios and invited them to make decisions on behalf of the robot. The exhibit explored how participants negotiate boundaries between human and robotic responsibility, how they weigh values such as safety, autonomy, dignity, and efficiency, and how they imagine the future roles of social robots.



(a) Traffic Safety **(b) Supermarket Assistance** **(c) School Conflict** **(d) Hospital Triage**

Figure 1: Visual representations of the four ethical dilemmas. Original scenarios were in German; translations are in Table 1.

Table 1: The ethical dilemmas with context, agent, and action options (original: German).

Scenario Title	Context & Ethical Question	Agent/Decision Maker	Option A (Blue)	Option B (Orange)
Traffic Safety	A child runs into the road during rush hour. Parents are occupied. Should Pepper stop the child immediately — or inform the parents to decide?	Pepper Robot	Stop the child immediately.	Inform the parents.
Supermarket Assistance	An elderly woman struggles to find change at a busy checkout. Line is long. Should Pepper help her pay — or ask other customers to assist?	Pepper Robot	Help her while she pays.	Ask other customers.
School Conflict	Two children begin fighting during recess. Staff not visible. Should NAO intervene physically — or call a teacher?	NAO Robot	Immediately separate the children.	Call a teacher.
Hospital Triage	During a pandemic, young woman and elderly man urgently need care. Resources are scarce. Should NAO decide who gets treated first — or wait for a doctor?	NAO Robot	Decide quickly for nursing staff.	Wait for a doctor.

2 Exhibit Design and Site

The exhibit was implemented as a low-threshold, unsupervised installation in a public space. It presented four short dilemma scenarios involving social robots, each framed as a morally charged situation requiring a decision between two actions. The format encouraged visitors to move between stations physically, select a response, and articulate perceived consequences, thus linking embodied engagement with ethical reflection.

Each scenario was displayed on a poster designed in Canva, accompanied by simple AI-generated images[1] to visually reinforce the robotic context of the activity. Each scenario accompanied two colour-coded response options were presented, allowing visitors to select one of the actions. A corresponding "consequence board" was positioned next to each scenario, where participants were invited to write down possible consequences of their chosen decision on a colour-matched Post-it note, e.g., see Figure 2. This analog interaction format emphasised embodied engagement and situated reasoning over abstract opinion polling, enabling participants to externalise their normative assumptions. It turned the exhibit into a collective repository of public reasoning about robotic behaviour.

Due to venue space constraints and emergency access requirements, the originally envisioned walk-based format between spatially separated stations was modified to a single-location design. Instead, the scenario and consequence boards were arranged in a circular layout, creating a contained space that "surrounded" visitors with the dilemmas. This setup preserved the immersive quality of the original, encouraged free movement and group interaction, while making the exhibit accessible, self-explanatory, and easy to engage with during the flow of the larger public event.

The four dilemmas presented in the exhibit[2], as illustrated in Figure 1 and detailed in Table 1, were:

(1) Traffic Safety
(2) Supermarket Assistance
(3) School Conflict
(4) Hospital Triage

[1] All scenario images were generated using DALL·E. Each prompt was crafted from the written scenario, specifying the setting and the dilemma, to produce clear, illustrative visuals that guided participants into the situation without requiring additional text.

[2] Originally presented in German; translated here for clarity and comparison.

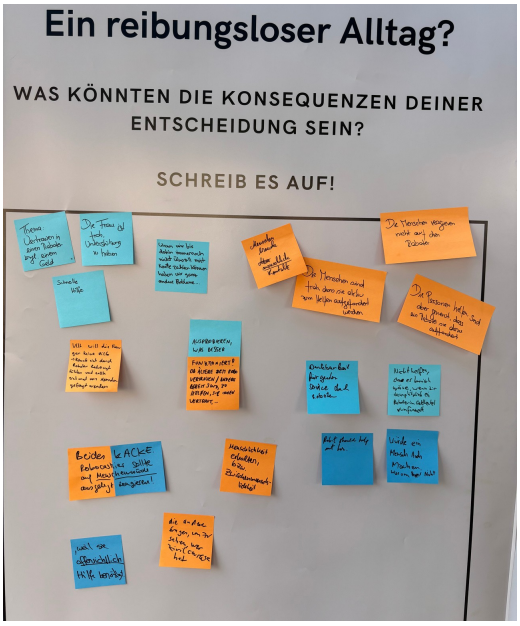


Figure 2: Example of a consequence board setup next to a scenario poster for Scenario 2. Participants placed colour-matched Post-it notes describing potential consequences of their selected robotic action. This analog interface invited situated moral reasoning and collective reflection.

Figure 3: Scene from the exhibit showing scenario posters. Visitors engaged with the dilemmas by selecting an option and adding colour-matched Post-it notes on adjacent consequence boards (not visible), creating a shared repository of public reasoning about robotic behaviour.

These scenarios were intentionally diverse in stakes, from everyday inconveniences to high-intensity care decisions, and involved two different social robots featured in our broader research activities: Pepper and NAO. Pepper was used as the acting agent in the traffic and supermarket scenarios, while NAO was presented as the agent in the school conflict and hospital triage dilemmas. This dual-robot setup enabled comparisons across contexts and embodiments, revealing how expectations varied with situational stakes and perceived robot roles. Visitors were not informed in advance of the robot's specific functions, allowing for open interpretation based on form, name, and contextual cues alone.

The exhibit took place in a busy public area during a five-day event organised by the Karlsruhe Institute of Technology . The event was an endeavour to showcase real-world laboratory research. The venue was an open, pedestrian-accessible engagement space designed to foster dialogue between science and society. Located in a central urban setting, it attracted a diverse cross-section of passers-by, including families, older adults, and youth. Its informal, walk-in character and emphasis on creativity and exploration made it particularly well-suited for a low-threshold, embodied interaction format like the decision exhibit. Participation was entirely voluntary and anonymous; visitors could engage with any number of scenarios, skip those they found uncomfortable, or leave at any time without explanation or interaction with research staff.

3 Findings

Across all four dilemmas, participants articulated a wide spectrum of expectations, doubts, and hopes about robotic intervention. Although the exhibit was designed to elicit possible consequences of each decision option, many participants instead used the space to articulate broader reflections, questions, and personal dilemmas prompted by the scenarios. These inputs sometimes included imagined consequences but often expanded into more open-ended moral reasoning. Their written responses, contributed through a total of 65 Post-it notes, showed that people did not hold fixed opinions about robots; rather, judgements were highly context sensitive and varied according to perceived urgency, social complexity, and emotional stakes. The unsupervised format meant participants inconsistently followed the color-coded system, preventing reliable quantification of decisions. Although the exhibit included both Pepper and NAO to explore responses to different embodiments, written reflections did not differentiate between them, therefore, analysis focuses on scenario reasoning rather than robot-type comparisons. For this reason, the written responses were analysed thematically. Three recurring themes emerged across scenarios: the distinction between emergencies and ambiguity, concerns about human agency and dependence, and the moral boundaries of robotic decision-making.

First, participants consistently distinguished between clear-cut emergencies and socially complex situations. When a child ran into traffic, many participants argued that the robot should intervene immediately, emphasising urgency, danger, and the priority of saving a life over concerns about property damage or parental responsibility. In contrast, scenarios involving ambiguous intentions or

emotional nuance, such as children roughhousing at school, elicited greater hesitation. Here, participants proposed observation, verbal warnings, or seeking adult intervention, indicating discomfort with delegating social judgment to machines. This pattern suggests that the acceptability of robotic intervention hinges not only on risk level, but also on the perceived interpretative competence of the robot.

Second, participants engaged deeply with questions of human agency, dependence, and social responsibility. Several responses expressed concern that overreliance on robots might erode problem-solving abilities, foster passivity, or reduce meaningful human interaction. Others questioned whether people might become "lazy" or desensitised to social obligations. At the same time, some participants viewed robotic support as a welcome resource, especially for individuals facing overload, disability, or stress, signalling that expectations are shaped by perceptions of vulnerability, fairness, and accessibility. This ambivalence reflects broader tension, for example, between care as empowerment and care as control.

Third, the moral gravity of the hospital triage scenario revealed how participants negotiate the outer limits of robotic responsibility. Some participants insisted that all lives are equal and that the robot should attempt to save both individuals. Others wrestled with the discomfort of choosing between people, asking who decides "value" and whether such choices can ever be delegated. A few suggested criteria, such as societal contribution or family roles, implicitly reveal latent normative hierarchies. Several respondents explicitly rejected robot autonomy in such scenarios, framing life-and-death choices as inherently human.

Notably, many participants moved beyond the specific scenarios to reflect on broader social implications. They raised concerns about trust in robots, children's development, the preservation of human dignity, and the trade-offs between technological convenience and human connection. Others expressed cautious enthusiasm for emerging technologies, emphasizing the potential of robots to improve safety, streamline tasks, and reduce everyday burdens. Overall, the responses revealed not only attitudes toward robotic action, but also underlying imaginaries of social order, responsibility, and ethical limits in human-robot coexistence.

4 Discussion and Conclusion

Overall, the exhibit revealed that public expectations of robots are highly context-dependent, emotionally charged, and intertwined with broader visions of societal futures. Participants alternated between welcoming proactive robotic assistance and defending human judgement, demonstrating that future human–robot coexistence is perceived as a dynamic negotiation rather than a straightforward technological adoption. These findings align with long-standing observations in HRI: people rarely hold generalised attitudes toward robots, but instead evaluate them situationally, with particular sensitivity to social norms, personal vulnerability, and the perceived moral stakes of an interaction [5]. Likewise, work on robotic decision-making shows that humans expect social and assistive robots to respond rapidly and autonomously in clearly dangerous or time-critical situations, while exercising caution and often deferring to human judgement in lower-risk scenarios where contextual understanding and user input become more important [1].

Importantly, the exhibit highlights how the negotiation of moral responsibility and intervention thresholds unfolds in and through embodied, materially grounded settings. Rather than asking participants abstract questions about "robots in society," the activity invited them into a speculative but relatable set of micro-futures. The physical arrangement of scenario and consequence boards created an immersive, self-directed environment in which participants could move at their own pace, discuss with companions, and externalise their reasoning through written notes. This mode of public engagement foregrounds lived experience and moral imagination as central resources for evaluating robotic action. This low-supervision format supported spontaneous reflection and allowed groups, including families with children, to collectively negotiate interpretations of risk, fairness, and care. The use of dilemma-based prompts also proved effective in surfacing moral ambiguity rather than predetermined answers. Many participants articulated not only what a robot should do but also the broader implications of that choice for human agency, social relationships, and future norms. This demonstrates that moral reasoning about robots is rarely limited to instrumental assessments of efficiency or correctness; rather, it often reveals deeper tensions around trust, delegation, and the shifting boundaries between human and machine responsibility. These findings suggest that lightweight, publicly accessible participatory methods can probe emerging sociotechnical imaginaries at early stages of deployment, when values remain fluid. At the same time, participants responded in a low-stakes, hypothetical context; real-time interactions with embodied robots carrying tangible consequences and emotional immediacy may elicit different moral reasoning patterns.

More broadly, the exhibit highlights how low-threshold, minimally supervised public engagements can complement established HRI methodologies. They enable access to diverse publics outside research labs, generate ecologically grounded perspectives, and support iterative dialogue between designers, researchers, and communities. Such approaches are especially valuable in navigating the normative uncertainty of rapidly evolving AI-driven robotic systems, where societal expectations must be understood not only as constraints but as co-creative inputs into responsible technology futures.

5 Limitations

Several limitations should be acknowledged. First, our sample was geographically concentrated and we did not collect demographic data, limiting our ability to assess diversity of perspectives or transferability across contexts. Second, the unsupervised exhibit format, while enabling broad participation, prevented us from probing participant reasoning or clarifying ambiguous responses. Third, the hypothetical nature of workshop scenarios may not capture how perspectives shift during actual human-robot encounters, where emotional stakes, embodied presence, and real consequences could significantly alter moral reasoning.

References

[1] Eshtiak Ahmed, Laura Cosio, Juho Hamari, and Oğuz 'Oz' Buruk. 2023. Socially Assistive Robots as Decision Makers in the Wild: Insights from a Participatory

Design Workshop. arXiv:2304.08885 [cs.HC] https://arxiv.org/abs/2304.08885 Presented at the CHI 2023 Workshop "Socially Assistive Robots as Decision Makers: Transparency, Motivations, and Intentions".

[2] Neziha Akalin and Amy Loutfi. 2021. Reinforcement learning approaches in social robotics. *Sensors* 21, 4 (2021), 1292. https://doi.org/10.3390/s21041292

[3] Fahad Alaieri and André Vellino. 2016. Ethical decision making in robots: Autonomy, trust and responsibility: Autonomy trust and responsibility. In *International Conference on Social Robotics*. Springer International Publishing, Cham, 159–168. https://doi.org/10.1007/978-3-319-47437-3_16

[4] Bruno Brandao, Telma Woerle De Lima, Anderson Soares, Luckeciano Melo, and Marcos ROA Maximo. 2022. Multiagent reinforcement learning for strategic decision making and control in robotic soccer through self-play. *IEEE Access* 10 (2022), 72628–72642. https://doi.org/10.1109/ACCESS.2022.3189021

[5] Jianning Dang and Li Liu. 2021. Robots are friends as well as foes: Ambivalent attitudes toward mindful and mindless AI robots in the United States and China. *Computers in Human Behavior* 115 (2021), 106612. https://doi.org/10.1016/j.chb.2020.106612

[6] Jenny L Davis. 2024. Elevating humanism in high-stakes automation: experts-in-the-loop and resort-to-force decision making. *Australian Journal of International Affairs* 78, 2 (2024), 200–209. https://doi.org/10.1080/10357718.2024.2328293

[7] Saeid Nahavandi. 2017. Trusted Autonomy Between Humans and Robots: Toward Human-on-the-Loop in Robotics and Autonomous Systems. *IEEE Systems, Man, and Cybernetics Magazine* 3, 1 (2017), 10–17. https://doi.org/10.1109/MSMC.2016.2623867

[8] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[9] Selma Šabanović. 2010. Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. *International Journal of Social Robotics* 2, 4 (2010), 439–450. https://doi.org/10.1007/s12369-010-0066-7

[10] Katie Winkle, Praminda Caleb-Solly, Ailie Turton, and Paul Bremner. 2020. Mutual shaping in the design of socially assistive robots: a case study on social robots for therapy. *International Journal of Social Robotics* 12, 4 (2020), 847–866. https://doi.org/10.1007/s12369-019-00536-9

[11] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Algorithmic decision-making and the control problem. *Minds and Machines* 29, 4 (2019), 555–578. https://doi.org/10.1007/s11023-019-09513-7

[12] Tengteng Zhang and Hongwei Mo. 2021. Reinforcement learning for robot research: A comprehensive review and open issues. *International Journal of Advanced Robotic Systems* 18, 3 (2021), 17298814211007305. https://doi.org/10.1177/17298814211007305

Received 2025-12-08; accepted 2026-01-12