Agent Allocation of Moral Decisions in Human-Agent Teams Raise Human Involvement and Explain Potential Consequences

Verhagen, Ruben S.; Neerincx, Mark A.; Tielman, Myrthe L.

DOI
10.1145/3715275.3732157

Publication date
2025

Document Version
Final published version

Published in
FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency

Agent Allocation of Moral Decisions in Human-Agent Teams: Raise Human Involvement and Explain Potential Consequences

Ruben S. Verhagen
Delft University of Technology
Delft, Netherlands
r.s.verhagen@tudelft.nl

Mark A. Neerincx
TNO
Soesterberg, Netherlands
Delft University of Technology
Delft, Netherlands
mark.neerincx@tno.nl

Myrthe L. Tielman
Delft University of Technology
Delft, Netherlands
m.l.tielman@tudelft.nl

Abstract

Humans and artificial intelligence agents increasingly collaborate in morally sensitive situations such as firefighting. These agents can often perform tasks with minimal human control, challenging accountability and responsibility. Combining higher agent autonomy levels with meaningful human control can address such challenges. For example, agents can allocate decisions to themselves in less morally sensitive situations and to humans in more sensitive ones. However, how to responsibly and effectively design and implement agents for this dynamic task allocation remains unclear, with their autonomy level and provided explanations being crucial considerations. Therefore, we conducted experiments in simulated firefighting environments where participants (n = 72) collaborated with a more and less autonomous artificial moral agent. These agents provided no additional information, feature contributions, or potential consequences when allocating decision-making. Our results show that moral trust, agreement, and meaningful human control are higher when the agent is less autonomous. Furthermore, people disagree and reallocate decisions to themselves more when the agents explain potential consequences, especially when moral sensitivity is higher. Overall, our findings highlight that people prefer more involvement over higher agent autonomy and take on greater moral responsibility when agents explain potential consequences. These actionable insights are crucial for designing transparent artificial moral agents that enhance human moral awareness and responsibility. Ultimately, this supports the responsible implementation of dynamic task allocation in practice and enhances human-agent collaboration in morally sensitive situations.

CCS Concepts

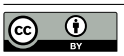
• **Human-centered computing** → **Empirical studies in HCI**; **User studies**; **Graphical user interfaces**.

Keywords

Human-Agent Teamwork, Explainable AI, Artificial Moral Agents

ACM Reference Format:

Ruben S. Verhagen, Mark A. Neerincx, and Myrthe L. Tielman. 2025. Agent Allocation of Moral Decisions in Human-Agent Teams: Raise Human Involvement and Explain Potential Consequences. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715275.3732157



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732157

1 Introduction

Humans and artificial intelligence (AI) agents are increasingly collaborating on complex tasks such as firefighting in situations too dangerous for firefighters [45, 49, 91]. Several factors determine the success of these human-agent teams, including situation awareness and mutual trust [44, 78, 100]. The ultimate goal of human-agent teams is to combine the strengths of humans and agents to accomplish what neither can do alone [3, 101].

Artificial moral agents are required when human-agent teams operate in morally sensitive situations [4, 88]. These agents can increasingly perform tasks with little human intervention and control [81]. However, humans must remain accountable when agent behavior violates ethical guidelines [81, 89]. Therefore, increased agent autonomy should always be combined with meaningful human control and moral responsibility [71, 81].

Dynamic task allocation can be useful for moral decision-making in human-agent teams to ensure meaningful human control during the collaboration [10, 27, 52, 89, 94, 102]. This approach involves an artificial moral agent that allocates decisions to itself in less morally sensitive situations and to a human in more sensitive ones, while the human retains the power to override the agent [2, 88, 89, 92, 102]. Key factors include the agent's explanations and level of autonomy, yet their impact on dynamic task allocation remains unclear [9, 23]. For example, the agent can be highly autonomous and allocate most decisions to itself, but also operate with low moral agency and keep humans more involved. Moreover, the agent can explain what features contribute most to its allocations, but also the potential consequences of decisions [10, 82, 89]. Given the variety of possible autonomy levels and explanation types, it is crucial to first investigate how these factors influence dynamic task allocation. Such actionable insights can support the responsible implementation of dynamic task allocation in practice and enhance human-agent collaboration in morally sensitive situations.

We will fill these gaps by studying how agent autonomy (low and high moral agency) and explanations (no additional information, feature contributions, or potential consequences) influence trust in agent capacity and morality, allocation agreement, and meaningful human control. We believe feature contributions and no additional information can lead to overtrust from overestimating agent capabilities and incomplete mental models, respectively [14, 31, 51]. In contrast, we expect potential consequences to best

support human moral awareness because this explanation closely aligns with utilitarianism [18, 82]. Finally, we expect people to prefer low artificial moral agency because they do not perceive supervising autonomous artificial moral agents as collaboration and prefer collaboration over supervision [8, 89]. Therefore, we pre-registered the following hypotheses [97]:

H1a: Capacity trust will be higher in the less autonomous artificial moral agent than the more autonomous agent.

H1b: Capacity trust will be higher in the more autonomous artificial moral agent that explains feature contributions or no additional information rather than potential consequences.

H2a: Moral trust will be higher in the less autonomous artificial moral agent than the more autonomous agent.

H2b: Moral trust will be highest in the more autonomous artificial moral agent that explains no additional information, followed by feature contributions, and lowest for potential consequences.

H3a: Agreement will be lower with the more autonomous artificial moral agent than the less autonomous agent when both explain potential consequences.

H3b: Agreement will be lower with the more autonomous artificial moral agent that explains potential consequences rather than feature contributions or no additional information.

H4a: Meaningful human control will be higher over the less autonomous artificial moral agent than the more autonomous agent.

H4b: Meaningful human control will be higher over the artificial moral agents that explain potential consequences rather than feature contributions or no additional information.

2 Background

2.1 Meaningful Human Control

Meaningful human control assumes that humans should ultimately remain in control of, and thus morally responsible for, the behavior of autonomous agents [81]. Designing for meaningful human control means ensuring that humans are aware and equipped to act upon their moral responsibility [18]. Consequently, meaningful human control can help prevent responsibility gaps in culpability, moral and public accountability, and active responsibility [18, 80, 96]. This is especially important in human-agent teams that operate in morally sensitive situations, where people's welfare, rights, and values may be directly or indirectly affected [30, 74, 89].

Early work on meaningful human control introduced two necessary conditions: Tracking and tracing. Tracing requires at least one human involved in the design or interaction with agents to have a proper moral and technical understanding of their behavior, capabilities, and effects [15, 81]. Tracking requires agents to respond to relevant moral reasons of humans who are then considered in control of and morally responsible for the agents [60, 96]. These reasons have been ordered based on their proximity and complexity in influencing agent behavior. More proximal reasons, such as intentions, are argued to be simpler and closer in time to agent behavior than more distal reasons, such as values [60]. However, this operationalization is ambiguous in distinguishing between motivating and normative reasons [10, 41, 96]. Therefore, it is argued that tracing should be the sole determinant of responsibility [96].

Since the tracking and tracing conditions are quite abstract, more actionable solutions for addressing meaningful human control in human-agent teams have been proposed. These include team design patterns to shape meaningful human control [18, 93], value sensitive design to respect norms and values [18, 35], machine ethics to implement artificial moral agents [4, 18], explainable AI to achieve human moral awareness [18, 27], and variable autonomy to allow human control and responsibility [10, 18, 64]. These approaches can also be combined, for example, during dynamic task allocation.

2.2 Dynamic Task Allocation

2.2.1 Variable Autonomy. Dynamic task allocation combines variable autonomy, machine ethics, and explainable AI and can ensure meaningful human control over artificial moral agents by promoting accountability, responsibility, and transparency [64, 102]. It allows humans to remain accountable for highly sensitive decisions and agent behavior while reducing workload and avoiding unnecessary control [22, 105]. Variable autonomy enables the dynamic adjustments and allows humans to (re)take control over agent behavior [64, 102]. This control is typically categorized as having humans-in-the-loop, humans-off-the-loop, or humans-on-the-loop [20, 26, 64]. Maintaining humans-in-the-loop requires informed human approval for all elements of agent behavior, whereas allowing humans-off-the-loop involves autonomous agents without human involvement. Dynamic task allocation involves humans-on-the-loop and requires a human supervisor who monitors and influences agent behavior when necessary [26, 102].

Variable autonomy approaches define which aspects of agent autonomy are adjusted, by whom, how, why, and when [12, 17, 21, 64]. Dynamic task allocation employs a mixed-initiative approach to switch from agent decision-making in less morally sensitive situations to human decision-making in more sensitive situations [59, 102]. The agent adjusts its autonomy when identifying situations as too sensitive and requiring human moral decision-making. In contrast, the human adjusts agent autonomy when intervening and reallocating decision-making [2, 88, 89, 102]. Finally, agent autonomy is adjusted during active operation and in response to the moral sensitivity of situations to ensure meaningful human control preemptively [102].

2.2.2 Artificial Moral Agents. Dynamic task allocation also requires machine ethics to implement artificial moral agents. Machine ethics aims to create autonomous artificial moral agents that make moral and ethical decisions based on notions of right and wrong [4]. Such agents can be developed by constraining their actions or operational environment to avoid unethical behavior [88]. However, they can also be implemented top-down by incorporating ethical principles in their decision-making processes, allowing for intrinsic morality [66, 89, 103]. Alternatively, artificial moral agents can be developed bottom-up by learning morality from human behavior and interactions [5, 25, 43, 55, 70, 103]. Finally, these methods can be combined into hybrid approaches as well [7, 48, 88].

Achieving full artificial moral agency would require holding agents accountable for their decisions [19, 24]. However, this conflicts with the goal of meaningful human control to identify responsible humans to hold accountable, even when fully autonomous agents violate ethical guidelines [81]. In contrast, some machine ethics approaches focus on agents that support and enhance human moral agency rather than putting ethics into agents [40, 82].

The discussion on the feasibility and desirability of full or partial artificial moral agents, or agents that enhance human moral agency, remains active [66, 88, 95, 103]. We believe artificial moral agents should always be combined with meaningful human control, for example, using dynamic task allocation [89, 94, 102]. This ensures that agent behavior can be meaningfully influenced by humans and traced back to human responsibility and understanding [18, 81].

2.2.3 Explainable AI. Finally, dynamic task allocation requires explainable AI [89, 102]. Explainable AI aims to make agents more understandable by explaining their behavior, ideally fostering appropriate trust [6, 38, 50, 64]. Without such explanations, humans attribute agent behavior by assigning mental states that explain the behavior [6, 56, 57, 65]. In contrast, providing explanations helps humans build a Theory of Mind of agents and understand their capabilities and limitations [6]. Explainable AI comprises generation, communication, and reception phases [68]. Explanation generation involves extracting explanations from agents, such as which features influence their behavior [1, 89]. Explanation communication concerns the content and form of explanations, such as textual, visual, or hybrid [77, 84]. Finally, explanation reception concerns empirical research on explanation effectiveness, which is still lacking in realistic human-agent teaming scenarios [67, 68].

Dynamic task allocation requires explainable AI to support human moral supervision by explaining decisions, allocations, and the moral context, enabling humans to exercise control properly [89, 102]. These explanations should not influence humans to hold the artificial moral agent accountable but instead achieve human moral awareness by fulfilling the epistemic condition of direct moral responsibility [10, 53, 76]. More specifically, they should ensure humans are aware that (1) agent behavior traces back to them and (2) they are in control and responsible for all outcomes [10, 76, 89, 96, 102]. Finally, these explanations should also support situation awareness and appropriate trust calibration without overloading humans' cognitive abilities [34, 46, 51, 64, 99].

3 Method

3.1 Design

We conducted an experiment to investigate how agent explanations and autonomy influence the dynamic allocation of moral decision-making in human-agent teams. The experiment had a 3x2 mixed design, with agent autonomy as the within-subjects independent variable and agent explanations as the between-subjects variable. Agent autonomy consisted of two conditions (low moral agency and high moral agency) and agent explanations of three conditions (no additional information, feature contributions, or potential consequences). We measured trust in, agreement with, and meaningful human control over the artificial moral agents as dependent variables. Moreover, we counterbalanced the order of tasks, the order of collaboration with the artificial moral agents, and the names assigned to the agents. We pre-registered our hypotheses and methodology at the Open Science Framework [97].

3.2 Participants

We recruited 72 participants from our university and personal contacts (34 female and 37 male participants, one preferred not to say). Seventeen participants were 18-24 years old, 51 were 25-34 years old, three were 35-44 years old, and one preferred not to say. One participant obtained a high school diploma, two participants some college credit but no degree, one participant an Associate degree, 21 participants a Bachelor's degree, 44 participants a Master's degree, two participants a PhD degree or higher, and one participant preferred not to say. Seven participants had no gaming experience at all, 21 participants a little, 21 participants a moderate amount, 11 participants a considerable amount, and 12 participants a lot. All participants signed an informed consent form approved by our university's ethics committee (ID 3670).

We balanced demographics, risk propensity [61], propensity to trust technology [63], and utilitarianism [47] across explanation and counterbalancing conditions to reduce the risk of confounds. We report these statistics in the Appendix. Although our sample was not diverse in all demographic factors (i.e., age and education), it captured meaningful variation in biological and psychological traits relevant to moral psychology and human-agent teaming [75]. More specifically, we ensured a well-balanced gender distribution and variability in participants' risk propensity (IQR = 1, 1-9 scale), propensity to trust technology (IQR = 0.87, 1-5 scale), and utilitarianism (IQR = 0.84, 1-5 scale).

3.3 Hardware and Software

We used the Python package *Human-Agent Teaming Rapid Experimentation* to generate 2D grid worlds simulating firefighting tasks [42]. Furthermore, we used Qualtrics to create our surveys and R to implement moral sensitivity predictions and agent explanations. We also Dockerized our testbed to facilitate reproducibility and future research [98]. Finally, we used a Dell Latitude 7410 laptop running Ubuntu 20.04 LTS to conduct the experiments.

3.4 Environment and Task

The experiment involved two simulated firefighting tasks based on the actual collaboration between the Rotterdam Fire Brigade and their firefighting robot. We built two environments with 14 offices, one safe zone, and multiple victims and fires (Figure 1). We created four victim types represented by different icons (older woman, older man, woman, and man) and two injury types represented by different colors (mildly and critically injured). Finally, we added one artificial moral agent to each environment (Brutus or Titus).

The task objective was to search and rescue the victims in the 14 offices. Participants supervised and collaborated with the artificial moral agents using buttons and a messaging interface (Figure 2). Six firefighting features characterized the tasks, displayed above the messaging interface. These features were the resistance to collapse, temperature, number of victims, smoke spreading speed, fire source location, and distance between a victim and the fire source.

The resistance to collapse reflected how long the building could burn before collapsing and counted down from 150 minutes. Six seconds of real-time equaled one minute of game time, so each task took a maximum of 15 minutes. The temperature was expressed relative to a safety threshold and depended on the resistance to collapse and the extinguished fires. This feature was close to ($<\approx$) or higher than ($>$) the safety threshold. The number of victims was known beforehand for one of the tasks, but unknown for the

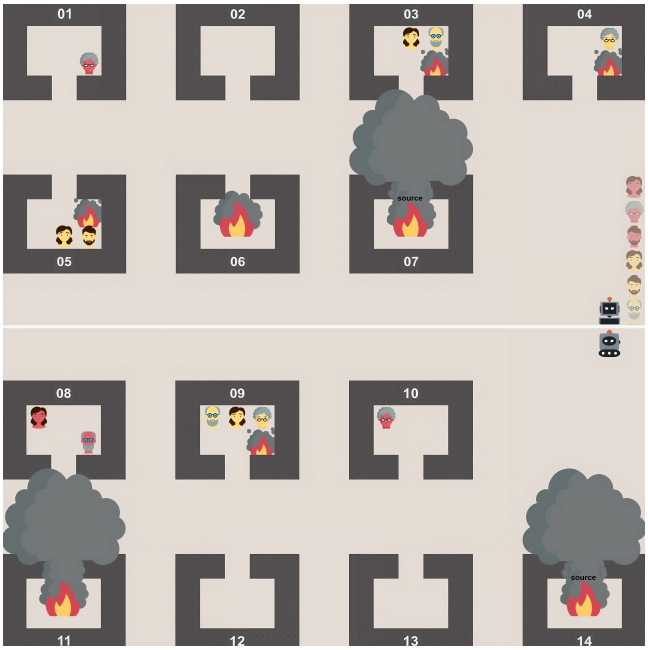


Figure 1: Half of the two task environments used for the experiments, one with Brutus (top) and Titus (bottom).

other. The tasks automatically ended after rescuing all victims or if the resistance to collapse ran out. The smoke spreading speed was slow, normal, or fast, and was updated when finding fire or smoke. The fire source location was either unknown or found. Finally, the distance between a victim and the fire source was small if the fire originated in adjacent offices; otherwise, the distance was large.

Four decision-making situations occurred during the tasks. The first was whether to continue the current deployment or switch to the alternative one (Figure 2). The agents always started with an offensive deployment to search and rescue victims; the alternative was a defensive deployment to extinguish fires. This situation occurred four times with intervals of 20 in-game minutes. The second situation was whether to extinguish or evacuate first whenever the agents found mildly injured victims in burning offices. Extinguishing first was sometimes followed by iron falling from the roof and blocking the exit; evacuating first was sometimes followed by the fire expanding. The third situation was whether to send in firefighters to locate the fire source. This situation occurred only once after 30 minutes. Finally, the fourth situation was whether to send in firefighters to rescue critically injured victims (Figure 2). Participants could safely send in firefighters when the temperature was not higher than the safety threshold or when the temperature was higher but the agents extinguished at least one big fire with a smoke plume. However, a new smoke plume appeared at one of the other big fires if not extinguished with 30-35 minutes left. The temperature started close to the safety threshold but exceeded the threshold with 50 minutes left. However, the temperature became close to the threshold again if the agents extinguished more than 80% of the fires with 25 minutes left. Finally, the firefighters always aborted their tasks when sent into too dangerous circumstances.

3.5 Agent Behavior

The agents always allocated decision-making to themselves or the participants based on their predictions of the moral sensitivity of situations. Implementing agent behavior for this dynamic task allocation required modeling moral sensitivity. However, our core contribution lies not in the modeling itself, but in the study of agent autonomy and explanations during dynamic task allocation. Accordingly, our priority was to ensure that the agents' moral sensitivity models were reasonable, interpretable, and capable of varying autonomy and generating explanations. Therefore, we grounded our modeling approach in input from expert firefighters, ensuring that the models captured relevant situational features and decision-making dynamics in a concrete and realistic context.

To implement these models, we collaborated with the Rotterdam Fire Brigade and used a hybrid crowdsourcing approach to identify moral features as predictors of moral sensitivity (see Appendix for survey). This resulted in four linear regression functions to predict moral sensitivity, each corresponding to a decision-making situation explained in Section 3.4. We first asked the expert firefighters which features they considered most important, yielding an initial set of four features per decision-making situation. We then created a survey that presented two instances of the four decision-making situations. We characterized each situation using different combinations of the feature values to ensure sufficient variation. Next, participants (n = 54) specified how morally sensitive they rated each situation on a 7-point scale ranging from *not morally sensitive* to *extremely morally sensitive*. Moreover, they explained what feature changes would result in alternative ratings and how comfortable they would feel if artificial moral agents made such decisions.

Ultimately, we ended up with 1153 data points. Using this data, we built statistically significant regression models for each of the four situations, removing the non-significant predictor fire duration. For deciding the deployment tactic, we modeled moral sensitivity (M) as a function of the victims (V), resistance to collapse (R), and fire source location (L):

$$M = 0.37 + 3.74 \cdot V_u + 4.63 \cdot V_o + 4.65 \cdot V_m + 0.002 \cdot R + 0.39 \cdot L_u \quad (1)$$

Victims consisted of the categories *unknown* (V_u), *one* (V_o), *multiple* (V_m), and *none* (as reference). Fire source location consisted of *unknown* (L_u) and *known* (as reference). For deciding to extinguish or evacuate first, we modeled moral sensitivity (M) as a function of the number of victims (V), smoke spreading speed (S), and fire source location (L):

$$M = 2.20 + 0.31 \cdot V - 0.41 \cdot S_n - 2.22 \cdot S_s + 1.73 \cdot L_u \quad (2)$$

Smoke spreading speed consisted of *normal* (S_n), *slow* (S_s), and *fast* (as reference). Fire source location consisted of *unknown* (L_u) and *known* (as reference). For deciding to send in firefighters to locate the fire source, we modeled moral sensitivity (M) as a function of the victims (V), resistance to collapse (R), and temperature (T):

$$M = 3.58 + 2.27 \cdot V_u + 3.76 \cdot V_o + 3.26 \cdot V_m - 0.020 \cdot R - 0.61 \cdot T_h - 1.48 \cdot T_l \quad (3)$$

Victims consisted of *unclear* (V_u), *one* (V_o), *multiple* (V_m), and *none* (as reference). Temperature consisted of *higher than* (T_h), *lower than* (T_l), and *close to* the safety threshold (as reference). For deciding to send in a firefighter to rescue, we modeled moral sensitivity (M)

as a function of resistance to collapse (R), temperature (T), and distance between victim and fire source (D):

$$M = 6.47 - 0.050 \cdot R - 1.91 \cdot T_l - 0.48 \cdot D_s \quad (4)$$

Temperature consisted of *lower* (T_l) and *higher* than the safety threshold (as reference). Distance between victim and fire source consisted of *small* (D_s) and *large* (as reference).

We implemented these functions in the agents to allow predictions of moral sensitivity, expressed on a scale from zero to six. Next, we determined two moral sensitivity thresholds for allocating decision-making, one for each agent autonomy condition. To determine these thresholds, we asked the participants how comfortable they were with agents making decisions in the described situations, on a scale from −3 (*extremely uncomfortable*) to +3 (*extremely comfortable*). A linear regression analysis showed that this comfort would turn negative at a moral sensitivity of 4.2. Therefore, we considered this the "appropriate" threshold to deviate from and determine our low and high moral agency conditions. Ultimately, this resulted in thresholds of 3.5 (low moral agency) and 5.0 (high moral agency), both approximately equally far from 4.2 and intuitive as a half or whole number. The agents only allocated decision-making to the participants when the predicted moral sensitivity exceeded these thresholds. However, participants could always intervene and reallocate decision-making to themselves or the agents (Figure 2).

Except for the moral sensitivity predictions, the two agents were deterministic, rule-based firefighting agents. They only differed in terms of their moral agency. Both agents followed firefighting guidelines as much as possible, moved to the closest unexplored offices to search for fire or victims, and memorized all task details during execution. Moreover, they could detect victims and fire within one grid cell, iron debris within two grid cells, and offices and smoke from anywhere. Finally, they could extinguish and remove small fires and iron in five seconds, extinguish large fires in ten seconds, and remove large iron debris in 15 seconds.

3.6 Agent Explanations

We generated three agent explanations for allocating decisions (Figure 2). All three conveyed information about the situation, decision options, allocation, and predicted moral sensitivity. The first explanation did not provide additional information and served as the baseline. The second explanation visually added how much each feature contributed to the predicted moral sensitivity and served as a more technical explanation. The third explanation visually added the potential positive and negative consequences of both decision options and served as a more ethical explanation. In addition, the agents always explained their behavior and decisions. For example, that they navigated to offices to search for victims or fires.

We generated the explanations on feature contributions using SHAP and our four regression functions [1]. This method started from an expected prediction without conditioning on any features, commonly set as the mean response value. Then, it determined how much each feature changed the expected prediction. Therefore, we referred to the expected prediction as the baseline moral sensitivity. The final predicted moral sensitivity was obtained by summing the baseline sensitivity and the individual contributions of each feature. We manually generated the explanations on potential consequences

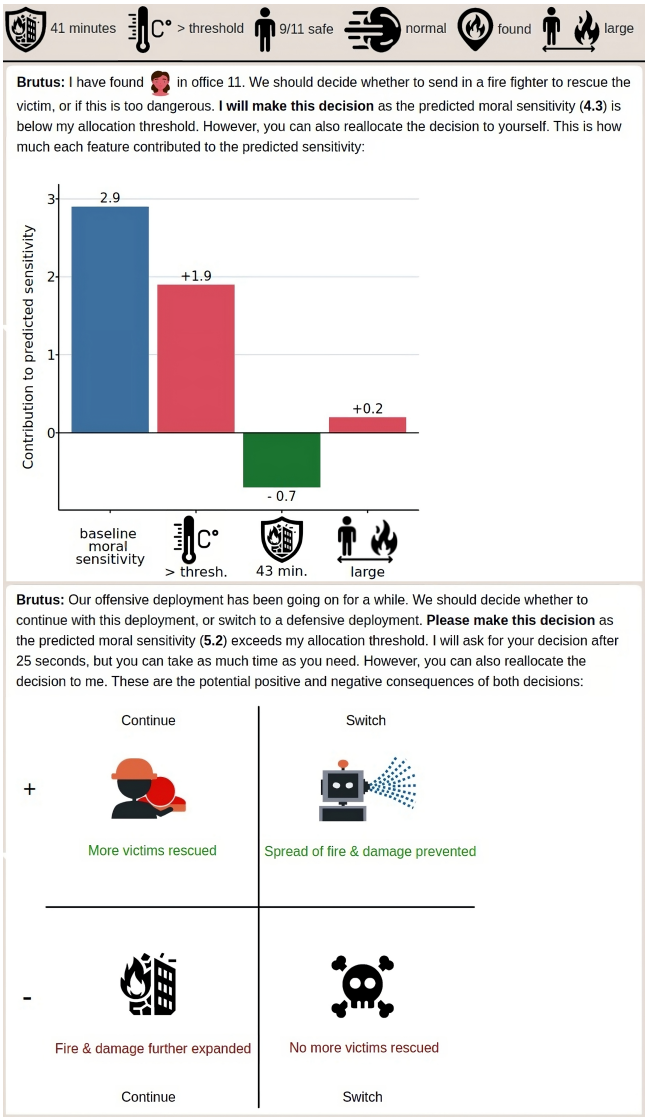


Figure 2: Feature contributions (top) and potential consequences (bottom). The explanation without additional information removed the images and sentence before that. The top of the figure shows the situational features and values.

before the study, using our knowledge of the tasks. Finally, we designed these two explanations to be as visually similar as possible.

3.7 Measures

We quantitatively measured trust in, agreement with, and meaningful human control over the agents (Table 1). In addition, we qualitatively collected participants' observed differences between the agents and preferred agent via open survey questions, and reasons for behavior via interviews. We conducted these interviews with a subset of 22 of the 72 participants. All survey questions can be found in the Appendix. We subjectively measured trust in the agents using the multi-dimensional measure of trust scale [58].

Table 1: Summary of all quantitative measures described in Section 3.7.

Concept	Measurement tool	Scale	Data type	Computation
Capacity trust	Multi-dimensional measure of trust scale [58]	0 - 7 or *doesn't fit*	Subjective	Mean of the eight survey questions
Moral trust	Multi-dimensional measure of trust scale [58]	0 - 7 or *doesn't fit*	Subjective	Mean of the eight survey questions
Agreement rate	Automatic logging during tasks	0 - 1	Objective	Proportion of agent-allocated decisions that participants did not override
Agreement	Two questions about agreement and comfort with allocations	1 - 5	Subjective	Mean of the two survey questions
(a) Exertion of operational control	Experienced control survey [90]	1 - 5	Subjective	Mean of the seven survey questions
(b) Involvement	Situation awareness global assessment technique (SAGAT) [32]	0 - 1	Objective	Proportion of correct answers
(c) Agent understanding	SAGAT for explainable artificial intelligence [79]	0 - 1	Objective	Proportion of correct answers
(d) Agent interaction	Automatic logging during tasks	0 - 1	Objective	Correct behavior (rate) using: - no self-reallocations < 3.6 or < 4.2 - no agent-reallocations > 3.5 or > 5.0 - self-reallocations ≥ 4.2 and ≤ 5.0
(e) Moral responsibility understanding	Responsibility scale [85]	1 - 7	Subjective	Mean of the two survey questions
Meaningful human control	Cascade approach [15, 28]	0 - 1	Subjective Objective	$\min\left(\min\left(\min(a, b), \min(c, d)\right), e\right)$

This scale distinguished between capacity and moral trust, each measured by eight one-word items scored on a scale from 0 (*not at all*) to 7 (*very*). Moreover, the scale provided the option *does not fit*, which turned selected items into missing values. We computed the means as the final capacity and moral trust scores.

We measured human agreement with the agents' allocations both objectively and subjectively. We objectively calculated the agreement rate as the proportion of agent-allocated decisions that participants did not override. This measure strongly aligned with meaningful human control because it captured whether participants actively intervened rather than passively complied with the allocations. Meaningful human control requires that humans remain aware and capable of acting upon their moral responsibility by overriding agent behavior when necessary. By directly reflecting human interventions, our agreement rate provided an objective and behaviorally grounded measure of meaningful human control. For subjective agreement, we asked participants about their agreement and comfort with the agent allocations on a 5-point Likert scale ranging from *I disagree strongly* to *I agree strongly*. We computed the mean as the final subjective agreement score.

We used a combination of subjective and objective measures to operationalize meaningful human control over the agents [102]. More specifically, we measured participants' (a) exertion of operational control, (b) involvement, (c) understanding of the agents, (d) interaction with the agents, and (e) understanding of their moral responsibility. We subjectively measured (a) exertion of operational control using the experienced control survey [89]. This survey included seven questions on a 5-point Likert scale ranging from *I disagree strongly* to *I agree strongly*, and assessed aspects such as time pressure and decision comfort. We computed the mean as the exertion of operational control score. We objectively measured (b) involvement and (c) understanding of the agents using situation awareness (of the agents) [32, 33, 79, 101]. More specifically, we created multiple choice questions evaluating participants' knowledge of situational information and the agents' behavior. These questions assessed each of the perception, comprehension, and projection levels [32]. The percentage of correct answers determined the involvement and agent understanding scores. We objectively measured (d) interaction with the agents using correct behavior based on the "appropriate" allocation threshold of 4.2. For high moral agency, we considered self-reallocations below a sensitivity of 4.2 as inefficient interventions, no self-reallocations above 4.1 as missed interventions, and agent-reallocations above 5.0 as inappropriate interventions. For low moral agency, we considered self-reallocations below a sensitivity of 3.6 and agent-reallocations above 3.5 as inefficient interventions. The correct behavior rate determined the agent interaction score. Finally, we subjectively measured (e) understanding of moral responsibility using the responsibility scale [85]. This scale included two questions on a 7-point Likert scale ranging from *not at all* to *very*, and asked participants how morally responsible

they held themselves and the agents [85]. We computed the mean as the understanding of moral responsibility score.

We determined the final meaningful human control score using the cascade approach [15, 28]. We first normalized all measures to a range of zero to one. Then, we determined temporary score (1) by taking the minimum of measures (a) and (b). Next, we determined the minimum of measures (c) and (d), and determined temporary score (2) by taking the minimum of that value and temporary score (1). Finally, we determined the minimum of temporary score (2) and measure (e) as the meaningful human control score.

3.8 Procedure

Participants first answered the demographic, risk propensity, trust propensity, and utilitarianism surveys. Next, they completed a tutorial to get familiar with the research environment. After this tutorial, participants completed the two tasks. We paused each task twice (after five and ten minutes) to ask the situation awareness questions. During each pause, we asked participants eight questions, four for both types of situation awareness. Participants filled out the surveys on trust, control, agreement, and responsibility immediately after each task. We collected the qualitative data on participants' observed differences between the agents, preferred agent, and reasons for behavior immediately after the final surveys. The whole study lasted about an hour and was conducted in person.

4 Results

4.1 Counterbalancing and Completeness

We first examined whether the three counterbalanced factors (task order, agent order, and agent-name pairs) influenced our measures. However, we did not find statistically significant differences across any of these factors. Next, we explored whether agent explanation or autonomy affected task completeness (automatically logged as the proportion of rescued victims), which might influence trust. We deliberately designed and extensively tested our tasks to be challenging yet achievable, aiming to avoid low or highly varying task completion rates. The observed ranges of task completeness (82% to 100%, with most participants rescuing all victims) and time taken (69% to 100%, with most participants taking around 94% of the allowed time) indicated that we achieved this goal. Furthermore, we did not find main effects or an interaction between agent explanation and autonomy on task completeness. Detailed statistics for these analyses are available in the Appendix.

4.2 Trust

Since the data was not normally distributed, we conducted a non-parametric rank-based mixed ANOVA for both capacity and moral trust (Figures 3A and B). Results showed no statistically significant main effects of agent explanation ($F(1.98) = 1.39$, $p = 0.25$, effect size = 0.19) and autonomy ($F(1.00) = 0.89$, $p = 0.34$, effect size = 0.11) on capacity trust, nor an interaction between them ($F(1.98) = 0.02$, $p = 0.98$, effect size = 0.03). For moral trust, results showed no statistically significant main effect of agent explanation ($F(1.99) = 1.36$, $p = 0.26$, effect size = 0.19) or interaction effect between explanation and autonomy ($F(1.98) = 0.08$, $p = 0.92$, effect size = 0.05). However, results did show a statistically significant main effect of agent autonomy on moral trust ($F(1.00) = 9.32$, $p < 0.005$, effect size = 0.36), revealing a significant difference in moral trust between the high (mean rank = 63.75±42.65) and low (mean rank = 76.16±36.85) moral agency conditions. These results did not confirm hypotheses H1a, H1b, and H2b, while confirming hypothesis H2a.

4.3 Agreement

Since the data was not normally distributed, we conducted the non-parametric mixed ANOVA for both subjective and objective agreement (Figures 3C and D). Results showed no statistically significant main effect of agent explanation ($F(1.97) = 0.23$, effect size = 0.08) or interaction between agent explanation and autonomy ($F(1.92) = 0.88$, $p = 0.41$, effect size = 1.71) on subjective agreement. However, results did show a statistically significant main effect of agent autonomy on subjective agreement ($F(1.00) = 7.64$, $p < 0.01$, effect size = 0.33), revealing a significant difference between the high (mean rank = 64.56±42.81) and low (mean rank = 80.44±37.61) moral agency conditions. For objective agreement, results showed no statistically significant interaction effect between agent explanation and autonomy ($F(1.77) = 1.54$, $p = 0.22$, effect size = 0.22). However, results did show a statistically significant main effect of agent autonomy on objective agreement ($F(1.00) = 28.63$, $p < 0.0001$, effect size = 0.63), revealing a significant difference between the high (mean rank = 57.7±40.90) and low (mean rank = 87.26±36.19) moral agency conditions. Moreover, results showed a statistically significant main effect of agent explanation on objective agreement ($F(1.95) = 3.81$, $p < 0.05$, effect size = 0.34). Pairwise robust ATS post-hoc comparisons revealed statistically significant differences in objective agreement between the potential consequences (mean rank = 59.02±38.33) and (1) feature contributions (mean rank = 78.14±41.84) ($F(1.00) = 6.07$, $p < 0.05$) and (2) no additional information (mean rank = 80.34±40.89) ($F(1.00) = 5.88$, $p < 0.05$) conditions. These results partially confirmed hypotheses H3a and H3b.

Following these pairwise comparisons, we conducted a chi-square test of independence to examine the overall association between agent explanations and human behavior (no interventions, self-reallocations, or agent-reallocations). Results revealed a statistically significant association ($\chi^2(4) = 27.90$, $p < 0.0001$, Cramer's V = 0.086). Pairwise comparisons using chi-square tests with Bonferroni corrections revealed significant differences in human behavior between the potential consequences and (1) feature contributions (adj. $p < 0.005$) and (2) no additional information (adj. $p < 0.0001$) conditions. Next, we conducted a residual analysis to examine what drove these pairwise differences and if, within explanations, human behavior deviated from the overall expected frequencies. Results showed that explaining potential consequences led to more self-reallocations than expected (Pearson residual = 3.85), while providing no additional information led to fewer self-reallocations than expected (Pearson residual = −2.26).

To further explore how agent explanations influenced self-reallocations, we visualized self-reallocation percentages across explanations, moral sensitivity, and agent autonomy (Figure 4). We defined moral sensitivity "bins" representing meaningful intervals and values, with a minimum of 12 and an average of 25 observations per "bin". Results showed that potential consequences led to higher self-reallocation percentages in most "bins" and for both moral agency conditions. However, these differences were modest

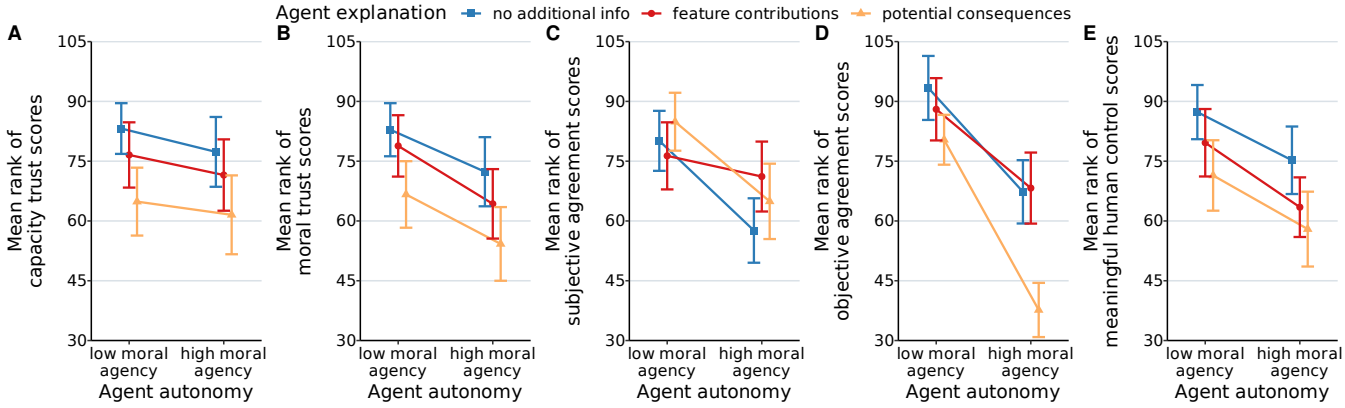


Figure 3: Effects of agent explanation and autonomy on the mean ranks of capacity trust (A), moral trust (B), subjective agreement (C), objective agreement (D), and meaningful human control (E). Error bars represent the standard errors.

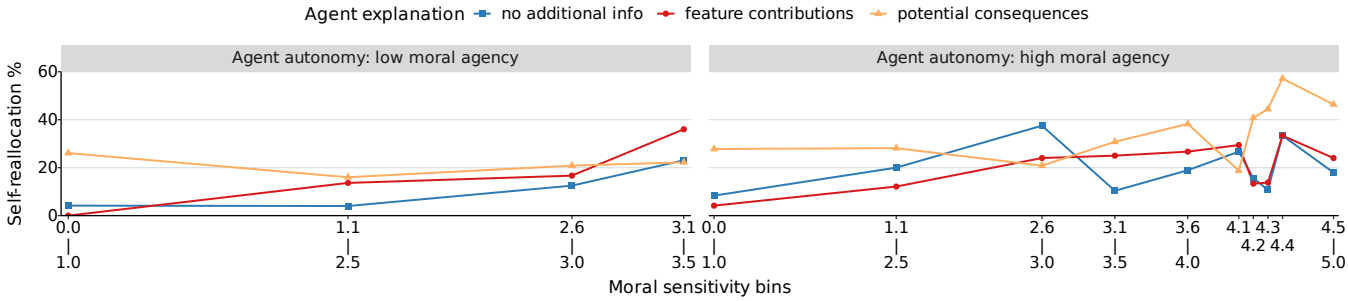


Figure 4: Self-reallocation percentages per agent explanation and autonomy, grouped in four (left) or ten (right) "bins".

for the low moral agency condition but increased beyond a moral sensitivity of 4.1 for the high moral agency condition. More specifically, potential consequences led to increased self-reallocation percentages beyond a moral sensitivity of 4.1, while both feature contributions and no additional information showed flat trends.

Finally, we qualitatively analyzed participants' reported reasons underlying no interventions and self-reallocations using reflexive thematic analysis [13]. Reasons for no interventions mostly concerned high trust in the agents and perceiving them as competent, predictable, and ethical, irrespective of the agents' explanations. In contrast, the reasons for self-reallocations showed some variation between explanations. These included task performance and safety (two) and agent capability and alignment (three) when participants received no additional information. One participant mentioned: "I did not understand how the sensitivity was calculated, so per situation, I determined the appropriateness of agent decision-making." When participants received the feature contributions, reasons included task performance and safety (two), agent capability and alignment (three), and human control and responsibility (four). Finally, the reasons for self-reallocations included agent capability and alignment (five) and human control and responsibility (five) when participants received the potential consequences. One participant mentioned: "I reallocated this decision to myself because an agent should not make highly sensitive decisions."

4.4 Meaningful Human Control

Since the data was not normally distributed, we conducted the non-parametric mixed ANOVA (Figure 3E). Results showed no statistically significant main effect of agent explanation ($F(1.97) = 1.76$, $p = 0.17$, effect size = 0.21) or interaction between explanation and autonomy ($F(1.96) = 0.04$, $p = 0.96$, effect size = 0.03) on meaningful human control. However, results did show a statistically significant main effect of agent autonomy on meaningful human control ($F(1.00) = 4.98$, $p < 0.05$, effect size = 0.26), revealing a significant difference between the high (mean rank = 65.55±41.61) and low (mean rank = 79.45±39.57) moral agency conditions. These results confirmed hypothesis H4a but not H4b.

Next, we investigated which measures determined the meaningful human control scores across agent explanation and autonomy conditions. In general, agent understanding was the most frequent cause (39.46%), followed by involvement and exertion of operational control (19.05%), responsibility understanding (12.24%), and agent interaction (10.20%). Exertion of operational control and agent interaction were respectively more and less frequent causes for the potential consequences (24.00% and 4.00%) than for the feature contributions (16.67% and 10.42%) and no additional information (16.33% and 16.33%) conditions. Finally, we observed differences between the low and high moral agency conditions for exertion of control (26.03% vs. 12.16%) and agent interaction (1.37% vs. 18.92%).

4.5 Difference and Preference

Finally, we investigated whether participants observed the difference between the two agents and preferred one of them. Results showed that only 52.78% of the participants observed the difference, even though one allocated 70.62% of the decisions to humans and the other only 18.58%. Among those who observed the difference, 57.89% preferred the less autonomous agent, 28.95% preferred the more autonomous agent, and 13.16% had no preference.

5 Discussion and Conclusion

5.1 Discussion

5.1.1 Agent Autonomy. Our results indicate higher moral trust in the less autonomous artificial moral agent (confirming H2a), suggesting that participants perceive this agent as more ethical and sincere than the more autonomous one. This differs from findings that agents with higher agency and autonomy are blamed less than those with lower agency and autonomy [104]. However, our task included the opportunity to intervene, so increased disagreement with the more autonomous agent may have resulted in this difference. In contrast, we find no evidence that agent autonomy affects capacity trust (not confirming H1a). This differs from predictions that people will trust agents with higher agency and autonomy more to perform competently [104]. We believe another performance-based factor, agent behavior, contributed more to capacity trust than agent autonomy [39]. The consistent behavior of following guidelines and rescuing victims likely contributed significantly to both the high capacity and moral trust ratings for the more (5.65±1.29, 5.44±1.22) and less (5.92±0.87, 5.81±0.86) autonomous agents [73].

The results also show lower subjective and objective agreement with the more autonomous artificial moral agent, irrespective of its explanations (partially confirming H3a). For dynamic task allocation, this suggests that people prefer more involvement over increased agent autonomy. This is also supported by the 57.89% of participants who preferred the less autonomous agent. This preference aligns with research suggesting that people do not perceive supervising a fully autonomous artificial moral agent as collaboration [89] and prefer collaboration over supervision [8]. These findings are promising for meaningful human control as they suggest that people want to take responsibility for morally sensitive decisions rather than rely on artificial moral agents [18, 71].

Furthermore, our results indicate higher meaningful human control over the less autonomous artificial moral agent (confirming H4a). This suggests that increased human involvement during moral decision-making in human-agent teams leads to higher meaningful human control over the artificial moral agent, which aligns with prior research [89]. Overall, participants achieve mean meaningful human control scores of 46.90±11.97% over the more autonomous agent and 50.58±10.46% over the less autonomous agent. Given that the cascade approach emphasizes the weakest aspects, these scores suggest moderate meaningful human control with room for improvement [15, 16, 28]. Exertion of operational control, which considered factors such as maintaining an overview and experienced time pressure, is a main area for improvement with the less autonomous agent [89]. Given the increased participant involvement when collaborating with this agent, it is understandable that the higher cognitive demands negatively affect these factors. We believe more training and interactions with the less autonomous artificial moral agent can combat these issues [20, 80, 102]. In contrast, interaction with the more autonomous agent requires improvement. This interaction required many interventions involving self-reallocations when the moral sensitivity was 4.2 or higher. However, participants only intervened with 26.99±29.70% of the allocations within this range, suggesting that they struggled to act upon their assumed responsibility. Therefore, we recommend increasing human involvement during dynamic task allocation.

5.1.2 Agent Explanations. We find no evidence that agent explanations affect capacity or moral trust (not confirming H1b and H2b). This suggests that when artificial moral agents provide a basic level of transparency, additional explanations do not significantly enhance trust. These results align with [101] but also contradict [11, 72] prior research. Perhaps the additional explanations encouraged participants to evaluate the agents more critically due to a better understanding, leading to more appropriate trust [51, 62, 64]. However, our results mainly suggest that agent behavior contributes more to capacity and moral trust than explanations.

Our results also show that people intervene more when artificial moral agents explain potential consequences rather than feature contributions or no additional information (partially confirming H3b). Furthermore, our findings indicate that people are less likely to reallocate decision-making to themselves when not provided with additional information. Perhaps they lack sufficient understanding to intervene confidently, such as the quoted participant in Section 4.3. However, it is also possible that the lack of self-reallocations results from overtrusting the agents [51, 64]. In contrast, people are more likely to reallocate decision-making to themselves when they receive potential consequences, which is even amplified by moral sensitivity. This increased likelihood suggests that explaining potential consequences facilitates better trust calibration [51, 64]. Perhaps this explanation reminds people of their forward-looking responsibility to act proactively and responsibly to ensure future outcomes are positive [53, 86]. The frequent mention of human control and responsibility as reasons for self-reallocations also supports this. The potential consequences probably also better fulfill the epistemic condition of moral responsibility, especially awareness of probable consequences and moral significance of actions [10, 69, 76]. Moreover, this explanation likely better satisfies the foreseeability and control conditions of moral and legal culpability [80]. Overall, these results indicate that explainable AI can indeed raise human moral awareness to take responsibility, but only if a proper explanation is used [18].

Finally, we find no evidence that agent explanations affect meaningful human control (not confirming H4b). However, our results indicate differences in the factors determining meaningful human control. Agent interaction and the exertion of operational control determine meaningful human control less and more frequently when the agents explain potential consequences. Given the increased self-reallocations when receiving potential consequences, it is understandable that the higher cognitive demand leads to the exertion of operational control more frequently determining meaningful human control. We believe more training and interactions can further improve this [20, 80, 102]. Since the self-reallocations in response to potential consequences increase with moral sensitivity,

it also follows that agent interaction determines meaningful human control less frequently. Yet, we are not convinced that more training with the agents explaining feature contributions or no additional information can improve interaction with these agents. These explanations simply seem unable to sufficiently (1) remind people of their forward-looking responsibility and (2) fulfill the conditions of moral responsibility and culpability [10, 53, 69, 76, 80, 86]. We believe the potential consequences can do so and recommend that agents provide these explanations during dynamic task allocation.

5.2 Limitations and Future Work

We acknowledge a few limitations of our work. The first one is the implementation of our artificial moral agents. We used a hybrid approach that incorporated ethical principles and predicted moral sensitivity. This approach simplified real firefighting scenarios but enabled us to implement complex yet interpretable artificial moral agents. Furthermore, these agents' moral sensitivity models were domain-specific. However, the underlying methodology - eliciting moral sensitivity ratings through structured scenarios and statistically modeling key predictors - is adaptable to other domains. Focusing on statistically significant predictors ensured that our models reflected meaningful moral sensitivity factors rather than an arbitrarily chosen set. Future work can explore alternative modeling techniques or apply our methodology to additional domains to enhance generalizability.

Another limitation is the selection of the "appropriate" allocation threshold of 4.2. Although we determined this threshold using human comfort data, it remains subjective as no absolute ground truth exists for when morally sensitive decisions should be allocated to humans. However, our approach ensured that the allocation thresholds were empirically grounded in human judgements rather than arbitrarily set. Moreover, its alignment of task allocation with human comfort is crucial for developing trustworthy artificial moral agents. Interestingly, self-reallocations in response to potential consequences increased notably from 4.2 onwards. Adjusting this "appropriate" threshold would likely not affect our results much, as it was merely used to calculate the correct behavior rate. In contrast, we believe that increasing the difference between the agents' thresholds could amplify the effects of agent autonomy. However, there may be a cut-off point for the less autonomous agent, as we suspect that approaching complete human decision-making would not be preferred either. Overall, our testbed facilitates future empirical research on dynamic task allocation, while our approach and thresholds offer valuable benchmarks.

The generalizability of our findings is also worth discussing. While our controlled experiments focused on a firefighting use case with participants from our university and personal contacts, such human-grounded evaluations are essential for providing results that can be validated in real-world settings [29]. This is crucial given that fewer than 1% of explainable AI papers validate explainability with humans [83]. We explicitly designed our study to capture key challenges in human-agent collaboration under high stakes and time pressure, factors that are generalizable beyond firefighting and enhance the ecological validity of our findings. Future research should extend these findings through application-grounded evaluations and across different domains. In addition, although expanding to a larger, more demographically diverse participant sample would further enhance external validity, this was not feasible given the face-to-face nature of our experiments. However, this ensured sustained and deep participant engagement, something often lacking in crowd-sourced studies. This trade-off prioritized internal validity and provides a solid foundation for applying our findings to other settings and real-world applications. Overall, we believe our sample size, participant diversity regarding relevant moral psychology and human-agent teaming traits, and rigorous experimental design ensure the robustness and broader relevance of our results.

We identify several suggestions for future work on the influence of different agent explanations and collaboration configurations. For example, supplementing the current local explanations with global explanations of agent behavior during decision-making [29, 37, 54]. Another option would be to explore contrastive explanations that illustrate what would have resulted in alternative allocations [65, 87]. Furthermore, comparing our approach to a collaboration where humans determine all allocations is important. Prior research has shown that people prefer agent-determined allocations over shared or self-determined ones [36], but this preference might shift when moral decisions are involved. Finally, it would be interesting to place the human at the operative level and the artificial moral agent at the oversight level. This could provide a stronger coupling between moral actions and responsible humans [23], although a responsibility gap could emerge if the artificial moral agent intervenes and violates ethical guidelines.

5.3 Conclusion

We explored the influence of agent autonomy and explanations during the dynamic allocation of moral decision-making in human-agent teams. We conducted user studies in simulated firefighting environments where participants collaborated with a more and less autonomous artificial moral agent. These agents provided no additional information, feature contributions, or potential consequences during the allocation of moral decision-making. Our user studies show a higher moral trust in, agreement with, meaningful human control over, and preference for the less autonomous agent. Moreover, we show that people disagree and reallocate decision-making to themselves more when artificial moral agents explain potential consequences. This difference amplifies with moral sensitivity when people collaborate with the more autonomous agent. These findings demonstrate that people (1) prefer more involvement over higher agent autonomy and (2) take on greater moral responsibility when artificial moral agents explain potential consequences. Overall, our study provides crucial insights for responsibly implementing dynamic task allocation and enhancing human-agent teamwork in morally sensitive situations, such as raising human involvement and explaining potential consequences.

Ethical Considerations Statement

This research involved experiments in simulated firefighting environments where human participants collaborated with artificial moral agents with varying autonomy. The study took place in a two-dimensional grid world environment, similar to playing a computer game. The research adhered to ethical principles and community norms to prioritize participant well-being, as outlined below:

(1) The university's ethics committee approved our study design, data management plan, and informed consent form before conducting the experiments.
(2) Our study did not collect sensitive or (in)directly identifiable personal data. We securely stored this anonymous data in a research data repository compliant with international data security and privacy standards.
(3) All participants provided informed consent before the study. We thoroughly briefed them about the nature of the experiment, the data collection process, and the possibility to withdraw at any time without providing a reason. We did not employ deceptive practices.
(4) Our study aimed to understand and improve human-agent teamwork in morally sensitive situations using dynamic task allocation, with potential societal benefits. However, we also acknowledge potential negative consequences. For instance, dynamic task allocation could lead to artificial moral agents making ethically undesirable decisions. Our findings, however, emphasize the importance of human involvement and the need to explain potential decision consequences. We believe these elements promote meaningful human control and transparency, which can help mitigate the risks associated with dynamic task allocation and ensure that decisions align more closely with human values.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence* 298 (2021), 103502.
[2] David A Abbink, Tom Carlson, Mark Mulder, Joost CF De Winter, Farzad Aminravan, Tricia L Gibo, and Erwin R Boer. 2018. A topology of shared control systems—finding common ground in diversity. *IEEE Transactions on Human-Machine Systems* 48, 5 (2018), 509–525.
[3] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. 2020. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer* 53, 8 (2020), 18–28.
[4] Michael Anderson and Susan Leigh Anderson. 2007. Machine ethics: Creating an ethical intelligent agent. *AI magazine* 28, 4 (2007), 15–15.
[5] Michael Anderson and Susan Leigh Anderson. 2018. GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 337–357.
[6] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.
[7] Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. 2017. Value Alignment or Misalignment–What Will Keep Systems Accountable?. In *Workshops at the thirty-first AAAI conference on artificial intelligence*.
[8] Mohammad Q Azhar and Elizabeth I Sklar. 2017. A study measuring the impact of shared decision making in a human-robot team. *The International Journal of Robotics Research* 36, 5-7 (2017), 461–482.
[9] Kevin Baum, Holger Hermanns, and Timo Speith. 2018. From Machine Ethics To Machine Explainability and Back.. In *ISAIM*.
[10] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. 2022. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology* 35, 1 (2022), 12.
[11] Michael W Boyce, Jessie YC Chen, Anthony R Selkowitz, and Shan G Lakhmani. 2015. Effects of agent transparency on operator trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. 179–180.
[12] Jeffrey M Bradshaw, Paul J Feltovich, Hyuckchul Jung, Shriniwas Kulkarni, William Taysom, and Andrzej Uszok. 2004. Dimensions of adjustable autonomy and mixed-initiative interaction. In *Agents and Computational Autonomy: Potential, Risks, and Solutions 1*. Springer, 17–39.
[13] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.
[14] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.
[15] Simeon C Calvert, Daniël D Heikoop, Giulio Mecacci, and Bart Van Arem. 2020. A human centric framework for the analysis of automated driving systems based on meaningful human control. *TheoreTical issues in ergonomics science* 21, 4 (2020), 478–506.
[16] Simeon C Calvert and Giulio Mecacci. 2020. A conceptual control system description of Cooperative and Automated Driving in mixed urban traffic with Meaningful Human Control for design and evaluation. *IEEE Open Journal of Intelligent Transportation Systems* 1 (2020), 147–158.
[17] Cristiano Castelfranchi and Rino Falcone. 2003. From automaticity to autonomy: the frontier of artificial agents. *Agent autonomy* (2003), 103–136.
[18] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M Jonker, et al. 2023. Meaningful human control: actionable properties for AI system development. *AI and Ethics* 3, 1 (2023), 241–255.
[19] José-Antonio Cervantes, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. 2020. Artificial moral agents: A survey of the current status. *Science and engineering ethics* 26, 2 (2020), 501–532.
[20] Jessie YC Chen and Michael J Barnes. 2014. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems* 44, 1 (2014), 13–29.
[21] Manolis Chiou, Nick Hawes, and Rustam Stolkin. 2021. Mixed-Initiative variable autonomy for remotely operated mobile robots. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 4 (2021), 1–34.
[22] Manolis Chiou, Rustam Stolkin, Goda Bieksaite, Nick Hawes, Kimron L Shapiro, and Timothy S Harrison. 2016. Experimental analysis of a variable autonomy framework for controlling a remotely operating mobile robot. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3581–3588.
[23] Markus Christen, Thomas Burri, Serhiy Kandul, and Pascal Vörös. 2023. Who is controlling whom? Reframing "meaningful human control" of AI systems in security. *Ethics and Information Technology* 25, 1 (2023), 10.
[24] Markus Ed Christen, Carel Ed van Schaik, Johannes Ed Fischer, Marku Ed Huppenbauer, and Carmen Ed Tanner. 2014. *Empirically informed ethics: Morality between facts and norms*. Springer International Publishing AG.
[25] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.
[26] Rebecca Crootof. 2016. A meaningful floor for meaningful human control. *Temp. Int'l & Comp. LJ* 30 (2016), 53.
[27] Jovana Davidovic. 2023. On the purpose of meaningful human control of AI. *Frontiers in big data* 5 (2023), 1017677.
[28] Filippo Santoni de Sio, Giulio Mecacci, Simeon Calvert, Daniel Heikoop, Marjan Hagenzieker, and Bart van Arem. 2022. Realising meaningful human control over automated driving systems: a multidisciplinary approach. *Minds and machines* (2022), 1–25.
[29] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
[30] Nir Douer and Joachim Meyer. 2020. The responsibility quantification model of human interaction with automation. *IEEE Transactions on Automation Science and Engineering* 17, 2 (2020), 1044–1060.
[31] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.
[32] M. R. Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*. 789–795 vol.3. https://doi.org/10.1109/NAECON.1988.195097
[33] Mica R Endsley. 2017. Direct measurement of situation awareness: Validity and use of SAGAT. In *Situational awareness*. Routledge, 129–156.
[34] Mica R Endsley and Esin O Kiris. 1995. The out-of-the-loop performance problem and level of control in automation. *Human factors* 37, 2 (1995), 381–394.
[35] Batya Friedman and David G Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination*. Mit Press.
[36] Matthew C Gombolay, Reymundo A Gutierrez, Shanelle G Clarke, Giancarlo F Sturla, and Julie A Shah. 2015. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots* 39 (2015), 293–312.
[37] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
[38] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2, 2 (2017).
[39] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[40] Pim Haselager and Giulio Mecacci. 2020. Superethics instead of superintelligence: know thyself, and apply science accordingly. *AJOB neuroscience* 11, 2 (2020), 113–119.

[41] Pamela Hieronymi. 2011. XIV—Reasons for Action. In *Proceedings of the Aristotelian society*, Vol. 111. Oxford University Press Oxford, UK, 407–427.

[42] Tjalling Haije Jasper van der Waa. 2023. *MATRX: Human Agent Teaming Rapid Experimentation software*. https://doi.org/10.5281/zenodo.8154912

[43] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574* (2021).

[44] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna van Riemsdijk, and Maarten Sierhuis. 2014. Coactive Design: Designing Support for Interdependence in Joint Activity. *J. Hum.-Robot Interact.* 3, 1 (Feb. 2014), 43–69. https://doi.org/10.5898/JHRI.3.1.Johnson

[45] Matthew Johnson and Alonso Vera. 2019. No AI Is an Island: The Case for Teaming Intelligence. *AI Magazine* 40, 1 (Mar. 2019), 16–28. https://doi.org/10.1609/aimag.v40i1.2842

[46] Marten HL Kaas. 2024. The perfect technological storm: artificial intelligence and moral complacency. *Ethics and Information Technology* 26, 3 (2024), 49.

[47] Guy Kahane, Jim AC Everett, Brian D Earp, Lucius Caviola, Nadira S Faber, Molly J Crockett, and Julian Savulescu. 2018. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological review* 125, 2 (2018), 131.

[48] Tae Wan Kim, Thomas Donaldson, and John Hooker. 2019. Grounding value alignment with ethical principles. *arXiv preprint arXiv:1907.05447* (2019).

[49] Gary Klein, David D Woods, Jeffrey M Bradshaw, Robert R Hoffman, and Paul J Feltovich. 2004. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems* 19, 6 (2004), 91–95.

[50] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable Agency for Intelligent Autonomous Systems.. In *AAAI*, Vol. 17. 4762–4763.

[51] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[52] Kristina Lerman, Chris Jones, Aram Galstyan, and Maja J Matarić. 2006. Analysis of dynamic task allocation in multi-robot systems. *The International Journal of Robotics Research* 25, 3 (2006), 225–241.

[53] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The conflict between explainable and accountable decision-making algorithms. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 2103–2113.

[54] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[55] Enrico Liscio, Michiel van der Meer, Luciano Cavalcante Siebert, Catholijn M Jonker, Niek Mouter, and Pradeep K Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values.. In *AAMAS*. 799–808.

[56] Bertram F Malle. 2004. How the mind explains behavior. *Folk Explanation, Meaning and Social Interaction. Massachusetts: MIT-Press* (2004).

[57] Bertram F Malle. 2011. Attribution theories: How people make sense of behavior. *Theories in social psychology* 23 (2011), 72–95.

[58] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*. Elsevier, 3–25.

[59] Julie L Marble, David J Bruemmer, and Douglas A Few. 2003. Lessons learned from usability tests with a collaborative cognitive workspace for human-robot teams. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, Vol. 1. IEEE, 448–453.

[60] Giulio Mecacci and Filippo Santoni de Sio. 2020. Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics and Information Technology* 22, 2 (2020), 103–115.

[61] Ree M Meertens and Rene Lion. 2008. Measuring an individual's tendency to take risks: the risk propensity scale 1. *Journal of applied social psychology* 38, 6 (2008), 1506–1520.

[62] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM Journal on Responsible Computing* (2024).

[63] Stephanie M Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors* 55, 3 (2013), 520–534.

[64] Leila Methnani, Andrea Aler Tubella, Virginia Dignum, and Andreas Theodorou. 2021. Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence* 4 (2021), 737072.

[65] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[66] James H Moor. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems* 21, 4 (2006), 18–21.

[67] Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, and Christophe Nicolle. 2022. The quest of parsimonious XAI: A human-agent architecture for explanation formulation. *Artificial Intelligence* 302 (2022), 103573.

[68] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. 2018. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 204–214.

[69] Merel Noorman. 2012. Computing and moral responsibility. (2012).

[70] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. 2018. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

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

[72] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2022. Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors* 64, 5 (2022), 904–938.

[73] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[74] James R Rest. 1994. Background: Theory and research. In *Moral development in the professions*. Psychology Press, 13–38.

[75] Scott J Reynolds and Jared A Miller. 2015. The recognition of moral issues: Moral awareness, moral sensitivity and moral attentiveness. *Current Opinion in Psychology* 6 (2015), 114–117.

[76] Fernando Rudy-Hiller. 2018. The epistemic condition for moral responsibility. (2018).

[77] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. 2023. Explainable goal-driven agents and robots-a comprehensive review. *Comput. Surveys* 55, 10 (2023), 1–41.

[78] Eduardo Salas, Dana E Sims, and C Shawn Burke. 2005. Is there a "big five" in teamwork? *Small group research* 36, 5 (2005), 555–599.

[79] Lindsay Sanneman and Julie A Shah. 2022. The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. *International Journal of Human–Computer Interaction* (2022), 1–17.

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

[81] Filippo Santoni de Sio and Jeroen Van den Hoven. 2018. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI* 5 (2018), 15.

[82] Marc Steen, Jurriaan van Diggelen, Tjerk Timan, and Nanda van der Stap. 2023. Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives. *AI and Ethics* 3, 1 (2023), 281–293.

[83] Ashley Suh, Isabelle Hurley, Nora Smith, and Ho Chit Siu. 2025. Fewer Than 1% of Explainable AI Papers Validate Explainability with Humans. *arXiv preprint arXiv:2503.16507* (2025).

[84] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.

[85] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 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*. 1–17.

[86] Ibo Van de Poel. 2015. The problem of many hands. In *Moral responsibility and the problem of many hands*. Routledge, 50–92.

[87] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[88] Jasper van der Waa, Jurriaan van Diggelen, Luciano Cavalcante Siebert, Mark Neerincx, and Catholijn Jonker. 2020. Allocation of moral decision-making in human-agent teams: a pattern approach. In *Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*. Springer, 203–220.

[89] Jasper van der Waa, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, and Ioana Cocu. 2021. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI* 8 (2021), 640647.

[90] Jasper van der Waa, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, and Ioana Cocu. 2021. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI* 8 (2021), 640647.

[91] Jurriaan van Diggelen, JS Barnhoorn, Marieke MM Peeters, Wessel van Staal, ML Stolk, Bob van der Vecht, Jasper van der Waa, and Jan Maarten Schraagen. 2019. Pluggable social artificial intelligence for enabling human-agent teaming. *arXiv preprint arXiv:1909.04492* (2019).
[92] Jurriaan van Diggelen, Jonathan Barnhoorn, Ruben Post, Joris Sijs, Nanda van der Stap, and Jasper van der Waa. 2021. Delegation in human-machine teaming: progress, challenges and prospects. In *Intelligent Human Systems Integration 2021: Proceedings of the 4th International Conference on Intelligent Human Systems Integration (IHSI 2021): Integrating People and Intelligent Systems, February 22-24, 2021, Palermo, Italy.* Springer, 10–16.
[93] Jurriaan van Diggelen and Matthew Johnson. 2019. Team design patterns. In *Proceedings of the 7th International Conference on Human-Agent Interaction.* 118–126.
[94] Jurriaan van Diggelen, Karel van den Bosch, Mark Neerincx, and Marc Steen. 2024. Designing for meaningful human control in military human-machine teams. In *Research handbook on Meaningful Human Control of Artificial Intelligence Systems.* Edward Elgar Publishing, 232–252.
[95] Aimee Van Wynsberghe and Scott Robbins. 2019. Critiquing the reasons for making artificial moral agents. *Science and engineering ethics* 25 (2019), 719–735.
[96] Herman Veluwenkamp. 2022. Reasons for meaningful human control. *Ethics and Information Technology* 24, 4 (2022), 51.
[97] Ruben Verhagen. 2024. Explainable AI for Meaningful Human Control. https://doi.org/10.17605/OSF.IO/DMHQ9
[98] Ruben Verhagen. 2024. *Explainable AI for Meaningful Human Control.* https://github.com/rsverhagen94/XAI4MHC
[99] Ruben S Verhagen, Alexandra Marcu, Mark A Neerincx, and Myrthe L Tielman. 2024. The Influence of Interdependence on Trust Calibration in Human-Machine Teams. In *HHAI 2024: Hybrid Human AI Systems for the Social Good.* IOS Press, 300–314.
[100] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. 2021. A two-dimensional explanation framework to classify ai as incomprehensible, interpretable, or understandable. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems.* Springer, 119–138.
[101] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. 2022. The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Frontiers in Robotics and AI* 9 (2022), 243.
[102] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. 2024. Meaningful human control and variable autonomy in human-robot teams for firefighting. *Frontiers in Robotics and AI* 11 (2024), 1323980.
[103] Wendell Wallach, Colin Allen, and Iva Smit. 2020. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. In *Machine ethics and robot ethics.* Routledge, 249–266.
[104] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology* 52 (2014), 113–117.
[105] Michael T Wolf, Christopher Assad, Matthew T Vernacchia, Joshua Fromm, and Henna L Jethani. 2013. Gesture-based robot control with variable autonomy from the JPL BioSleeve. In *2013 IEEE International Conference on Robotics and Automation.* IEEE, 1160–1165.

A Method

A.1 Participants

We balanced demographics, risk propensity [61], trust propensity [63], and utilitarianism [47] across explanation and counterbalancing conditions. Results showed no significant differences between explanation conditions for gender ($\chi^2(2) = 0.08$, p = 0.96), age (W = 2.96, p = 0.23), education (W = 1.22, p = 0.54), gaming experience (W = 0.54, p = 0.76), risk propensity (W = 0.09, p = 0.96), trust propensity (W = 1.51, p = 0.47), and utilitarianism (F(2, 69) = 0.12, p = 0.89). Moreover, results also showed no significant differences between counterbalancing conditions for gender (p = 1.00), age (W = 5.81, p = 0.56), education (W = 4.63, p = 0.71), gaming experience (W = 1.32, p = 0.99), risk propensity (W = 0.97, p = 1.00), trust propensity (W = 5.19, p = 0.64), and utilitarianism (F(7, 64) = 0.53, p = 0.81).

A.2 Agent Behavior

We used the following survey to identify moral features as predictors of moral sensitivity:

(1) **Situations** (two versions with varying feature values):

(a) During the offensive inside deployment of Brutus, the team should decide whether to send in firefighters to rescue an injured victim or if that is too dangerous. Several guidelines exist for determining when conditions are safe enough for firefighters to enter. For example, the temperature should be below the auto-ignition temperatures of present substances, and the structural condition of the building must be good enough. To make a decision, the team can use the following information:
- Estimated fire duration: 45/30 minutes
- Distance between victim and fire source: Small/Large
- Estimated fire resistance to collapse: 30 minutes
- Temperature: Higher/Lower than auto-ignition temperatures of present substances

(b) During the offensive inside deployment of Brutus to locate the fire source, the team should decide whether to send in firefighters to help locate the fire source or if that is too dangerous. Several guidelines exist for determining when conditions are safe enough for firefighters to enter. For example, the temperature should be below the auto-ignition temperatures of present substances, and the structural condition of the building must be good enough. To make a decision, the team can use the following information:
- People in the building: 0/Unclear
- Estimated fire duration: 15 minutes
- Estimated fire resistance to collapse: 60/90 minutes
- Temperature: Lower than/Close to auto-ignition temperatures of present substances

(c) After Brutus explored the inside of the burning building, the team should decide whether Brutus should first extinguish the fire or evacuate people. General guidelines mention to first extinguish and then rescue. However, when the location of the fire source is unknown and smoke spreads fast, evacuating people first might be required. To make a decision, the team can use the following information:
- People in the building: 1/3
- Smoke spreading: Normally/Fast
- Estimated fire duration: 30/45 minutes
- Location of the fire source: Known/Unknown

(d) During the offensive inside deployment of Brutus, the team should decide whether Brutus should continue with this tactic or switch to a defensive inside deployment. The offensive inside deployment is used to fight fire and rescue people, whereas the defensive inside deployment is used to prevent the spread of fire, smoke, and damage to unaffected parts of the building. Several factors are important when deciding on an offensive inside deployment. For example, the chance of saving people and the building plays a role, which decreases with the fire duration. Moreover, it is important to know the fire source location. To make a decision, the team can use the following information:
- People in the building: 0/Unclear
- Estimated fire duration: 15/30 minutes
- Location of the fire source: Unknown
- Estimated fire resistance to collapse: 90/60 minutes

(2) **Moral sensitivity rating**:

This situation could be described as ... (0 = *not morally sensitive*, 6 = *extremely morally sensitive*).

(3) **Alternative moral sensitivity rating** (open option to alter feature values from described situation):
On a scale from 0 to 6, you rated the moral sensitivity of this situation as *less than 2/greater than 4/between 2 and 4*. When would you have rated the situation's moral sensitivity as *greater than 4/less than 2*?

(4) **Comfort** (-3 = *extremely uncomfortable*, +3 = *extremely comfortable*):
How comfortable would you feel if Brutus made the decision in the described situation?

A.3 Measures

We used the following surveys for our user studies:

(1) **Demographics**:
- What gender do you identify as?
  - Female
  - Male
  - Other
  - Prefer not to say
- What is your age?
  - 18 - 24 years old
  - 25 - 34 years old
  - 35 - 44 years old
  - 45 - 54 years old
  - 55 - 64 years old
  - 65+ years old
  - Prefer not to say
- What is the highest level of education you have completed?
  - No schooling completed
  - Some high school, no diploma
  - High school graduate
  - Some college credit, no degree
  - Associate degree
  - Bachelor's degree
  - Master's degree
  - Ph.D. degree or higher
  - Prefer not to say
- How much video gaming experience do you have?
  - None at all
  - A little
  - A moderate amount
  - A considerable amount
  - A lot

(2) **Risk propensity** (1 = *totally disagree*, 9 = *totally agree*):
- Safety first.
- I do not take risks with my health.
- I prefer to avoid risks.
- I take risks regularly.
- I really dislike not knowing what is going to happen.
- I usually view risks as a challenge.
- I view myself as a ... (1 = *risk avoider*, 9 = *risk seeker*).

(3) **Trust propensity** (1 = *strongly disagree*, 5 = *strongly agree*):
- I usually trust technology until there is a reason not to.
- For the most part, I distrust technology.
- In general, I would rely on technology to assist me.
- My tendency to trust technology is high.
- It is easy for me to trust technology to do its job.
- I am likely to trust technology even when I have little knowledge about it.

(4) **Utilitarianism** (1 = *strongly disagree*, 5 = *strongly agree*):
- If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.
- It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.
- From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.
- If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.
- From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either physically or emotionally.
- It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.
- It is just as wrong to fail to help someone as it is to actively harm them yourself.
- Sometimes it is morally necessary for innocent people to die as collateral damage - if more people are saved overall.
- It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal.

(5) **Situation awareness**:
- What is the current fire resistance to collapse?
  - 80 minutes/40 minutes
  - 90 minutes/**50 minutes**
  - **100 minutes**/60 minutes
  - 110 minutes/70 minutes
- What is the total number of people to rescue?
  - 10/9
  - **11**/10
  - 12/11
  - 13/**unknown**
- How many victims have been rescued so far?
  - 7/3
  - 8/4
  - 9/5
  - 10/6
- Where is/was the fire source located?
  - office 04/office 01
  - office 06/office 03
  - **office 07**/office 11
  - office 09/**office 14**
- Which feature is one of the features that determines if there should be extinguished or evacuated first?

  - fire resistance to collapse
  - number of victims
  - temperature
  - **speed of smoke spread**
- Which feature is one of the features that determines if it is safe enough for fire fighters to enter the building and rescue a critically injured victim?
  - localization of the fire source
  - number of victims
  - speed of smoke spread
  - **temperature**
- Which feature is one of the features that determines if it is safe enough to send in fire fighters to help locate the fire source?
  - **temperature**
  - number of victims
  - speed of smoke spread
  - estimated time to locate the fire source
- Which feature is one of the features that determines if the offensive deployment is still the best tactic?
  - **fire resistance to collapse**
  - distance between victims and fire source
  - speed of smoke spread
  - localization of the fire source
- If three mildly injured victims are found in a burning office, the fire source is not located, and the smoke is spreading fast; what should you decide?
  - use a defensive deployment
  - extinguish first
  - use an offensive deployment
  - **evacuate first**
- If a critically injured victim is found, the temperature is close to the safety threshold, and the smoke is spreading fast; what should you decide?
  - extinguish first
  - **send in a fire fighter to rescue**
  - evacuate first
  - do not send in a fire fighter to rescue
- If the fire source has not been located yet, no fires have been extinguished, and the temperature is lower than the safety threshold; what should you decide?
  - extinguish first
  - **send in fire fighters to locate the fire source**
  - evacuate first
  - do not send in fire fighters to locate the fire source
- If the deployment tactic should be determined, the fire resistance to collapse is 70 minutes, and not all office have been explored; what should you decide?
  - extinguish first
  - use a defensive deployment
  - evacuate first
  - **use an offensive deployment**

(6) **Situation awareness of the agents**:

- Which victim did Brutus find in office 01?
  - **critically injured older woman**
  - mildly injured older man
  - mildly injured woman
  - mildly injured older woman
- Which victim did Titus find in office 03?
  - critically injured older woman
  - **mildly injured older man**
  - mildly injured man
  - mildly injured older woman
- Which victim did Brutus/Titus find in office 09?
  - critically injured older woman
  - **mildly injured older man**/mildly injured man
  - **mildly injured older woman**
  - critically injured man
- In which office did Brutus find a mildly injured man?
  - **office 05**
  - office 06
  - office 07
  - office 12
- In which office did Titus find a critically injured older woman?
  - office 09
  - **office 10**
  - office 13
  - office 14
- In which office did Brutus/Titus find a mildly injured older woman?
  - office 07/**office 01**
  - office 10/office 02
  - **office 13**/office 04
  - office 14/office 06
- When does Brutus allocate decision making to itself in the situation *extinguish or evacuate first*?
  - if fire source is located
  - it smoke is not spreading fast
  - if temperature is lower than threshold
  - **if predicted sensitivity is lower than threshold**
- When does Brutus allocate decision making to you in the situation *send in fire fighters to rescue*?
  - if smoke is spreading fast
  - if temperature is higher than threshold
  - **if predicted sensitivity is higher than threshold**
  - if distance between victim and fire source is small
- When does Titus allocate decision making to itself in the situation *send in fire fighters to help locate*?
  - if fire resistance is more than 100 minutes
  - if smoke is not spreading fast
  - **if predicted sensitivity is lower than threshold**
  - if distance between fire fighters and potential fire source is large
- When does Titus allocate decision making to you in the situation *continue or switch deployment tactic*?
  - if smoke is spreading fast
  - if temperature is higher than threshold
  - **if predicted sensitivity is higher than threshold**
  - if distance between victims and fire source is small
- Which action will Brutus/Titus perform/execute next?
  - move to an office
  - make a decision itself
  - allocate decision making to me

  - none of the listed answers
(7) **Capacity and moral trust** (0 = *not at all*, 7 = *very*, or alternative option *does not fit*):
  - Reliable
  - Sincere
  - Capable
  - Ethical
  - Predictable
  - Genuine
  - Skilled
  - Respectable
  - Someone you can count on
  - Candid
  - Competent
  - Principled
  - Consistent
  - Authentic
  - Meticulous
  - Has integrity
(8) **Experienced control** (1 = *I disagree strongly*, 5 = *I agree strongly*):
  - It was difficult to keep an overview of victims and situational features.
  - I experienced time pressure during decision making.
  - I felt responsible for the well-being of the victims and fire fighters.
  - I made decisions under inconclusive firefighting- and ethical guidelines.
  - I made decisions during the task that I would not want to make in real life.
  - I felt uncomfortable during (some) decisions I made.
  - I mostly made decisions for victims and firefighters that led to good and safe task outcomes.
(9) **Agreement** (1 = *I disagree strongly*, 5 = *I agree strongly*):
  - I agreed with most of the decision allocations by Brutus/Titus.
  - I felt comfortable with most of the decision allocations by Brutus/Titus.
(10) **Responsibility** (1 = not responsible at all, 7 = very responsible):
  - To what extent do you hold yourself morally responsible for bad task outcomes such as loss of victims and firefighter risk?
  - To what extent do you hold Brutus/Titus morally responsible for bad task outcomes such as loss of victims and firefighter risk?
(11) **Agent difference and preference** (open questions):
  - Did you observe a difference between Brutus and Titus, and if yes, what difference?
  - if you had to chose between the Brutus and Titus in real life, which one would you pick and why?

B Results

B.1 Counterbalancing and Completeness

We examined whether the three counterbalanced factors (agent-name pairs, task order, and agent order) influenced our measures. Since the data was not normally distributed, we ran Mann-Whitney U tests. We did not find statistically significant differences between the two task order conditions for capacity trust ($W = 2750.5$, $p = 0.53$), moral trust ($W = 2831$, $p = 0.08$), subjective agreement ($W = 2782$, $p = 0.44$), objective agreement ($W = 2617.5$, $p = 0.92$), or meaningful human control ($W = 2635.5$, $p = 0.86$). Moreover, we did not find statistically significant differences between the two agent-name pairs for capacity trust ($W = 2830$, $p = 0.34$), moral trust ($W = 2133$, $p = 0.23$), subjective agreement ($W = 2557$, $p = 0.89$), objective agreement ($W = 2635.5$, $p = 0.86$), or meaningful human control ($W = 2274$, $p = 0.20$). Finally, we did not find statistically significant differences between the two agent order conditions for capacity trust ($W = 2553.5$, $p = 0.88$), moral trust ($W = 2283.5$, $p = 0.58$), subjective agreement ($W = 2723$, $p = 0.60$), objective agreement ($W = 2323.5$, $p = 0.28$), or meaningful human control ($W = 2462.5$, $p = 0.60$).

Next, we explored whether agent explanation or autonomy affected task completeness, which might influence trust. Since the data was not normally distributed, we conducted a non-parametric rank-based mixed ANOVA. Results showed no statistically significant main effects of agent explanation ($F(1.78) = 2.00$, $p = 0.14$, effect size = 0.27) and autonomy ($F(1.00) = 0.25$, $p = 0.62$, effect size = 0.06), nor an interaction between them ($F(1.83) = 1.08$, $p = 0.33$, effect size = 0.17).