scientific reports

OPEN

The power of justifications to repair human-robot trust, even under moral disagreement

Elizabeth K. Phillips[1]✉ & Bertram F. Malle[2]

To avert criticism and losses of trust, robots that adopt social roles in the near future will have to be aware of and follow the norms of the communities in which they operate. However, norms often conflict with one another, and resolving such conflicts requires prioritizing one norm and violating the other, conflicting norm. As a result, robots will face moral disapproval from at least some of their human interaction partners. We investigate a powerful tool that humans use—and autonomous agents should use—to manage such moral disapproval and maintain trust: ***justifications***, which explain not just why the agent acted but what norms and values the action upheld. In three experiments (***N*** = 3,596), we demonstrate, replicate, and generalize that justifications, more than mere explanations, mitigate moral disapproval and recover robots' perceived trustworthiness, even when the robot's action is in direct moral disagreement with the human observer. We conclude that people simultaneously blame the robot for its specific norm-violating action and appreciate the robot's integrity to make trustworthy decisions. Justifying norm-violating actions may allow robots to become better integrated into human communities and adopt social roles that will involve morally significant decisions.

Keywords Justification, Human-robot trust repair, Moral judgment, Blame, Moral disagreement, Moral dilemmas

People are no longer asking if social robots will be part of human communities; they are wondering when robots will arrive, and in what roles and contexts. In fact, robots are beginning to appear not only as service personnel[1], soldiers[2], and astronauts[3], but also as caregivers[4], teachers[5], and even companions and partners[6–8]. Autonomous machine decision making in these roles will require a blend of risk awareness, social skills, and moral appropriateness.

Previous standards for machine decision making have focused on reliability—consistent, repeated performance under known conditions[9,10]. However, reliability alone will not suffice to meet the demands of complex social domains. When humans interact in these domains, they do more than strive for reliability; they strive for socially and morally appropriate behavior in line with the norms of their communities[11]. But should robots have these skills? And if so, how should we equip them?

Some scholars have warned against designing moral robots, or robots that have significant autonomy and responsibility in human affairs of moral consequence[12,13]. Healthy skepticism is important, and a conservative position would be viable if we could collectively decide to keep robots out of militaries, schools, hospitals, and private homes. But that option may no longer be available. Robots are already entering those domains, where they advise or decide on significant issues such as firing missiles[14] or providing triaging decisions when allocating medical care[15,16]. The best option is to make these robots as safe, beneficial, and socially and morally appropriate as possible, despite their obvious deficits[17].

Norm competence and norm conflict for moral machines

One pathway to morally appropriate robots lies in designing machines that have norm competence—the capacity to be aware of, follow, and prioritize the norms of the communities in which they operate[18]. A norm is an instruction to perform, in a given context, a particular action that other community members also perform and, importantly, demand of each other to perform[19–22]. Researchers from many disciplines have recognized the unique role that normative considerations play in human rational choice[23], economics[24–26], legal foundations[27,28], and in regulating institutions and cultures[29]. As such, norms are central tools of social influence and regulation in all human communities. They constrain individuals' self-interest in favor of the group's interest, they increase the mutual predictability of behavior for both individuals and groups, and, as a result, they can foster trust and

[1]George Mason University, Fairfax, V.A, USA. [2]Brown University, Providence, R.I, USA. ✉email: ephill3@gmu.edu

group cohesion. Thus, it would seem highly desirable if sophisticated, trustworthy artificial agents had norm competence.

But even if we succeed in implementing such norm competence in robots, a substantial challenge will inevitably arise: Norms often conflict with one another. Sometimes an answer to a person's question must either be polite and dishonest or honest and impolite; sometimes being fair requires breaking a friend's expectation of loyalty; sometimes only one of two candidates can receive a donor kidney. Recent literature has identified a number of potential norm conflicts for robots and other artificial agents—from self-driving cars to autonomous military drones, from home assistants to care robots[30–33]. For example, in space exploration, robots may make difficult choices between risks to material and mission focus; physical therapy robots may make choices between discomfort to a patient and effectiveness of an exercise. The more we see robots take part in human communities and take on significant social roles, the more they will face such norm conflicts. How they resolve them, and explain their decisions, will be critical for maintaining human trust in these machines and facilitating their long-term integration into human communities.

Resolving norm conflicts

The only way to resolve conflicts between norms (even just two) is by adhering to the norm one considers more important and violating the other, less important norm[34,35]. This implies that any norm conflict resolution will involve an inevitable norm violation, which can result in moral disapproval and loss of trust[36,37]. Therefore, even norm-competent robots will intentionally violate some norms some of the time. How can they handle the likely resulting moral criticism and loss of trust?

One possibility is to make robots more transparent[38–41]. However, transparency primarily combats the challenge of machine opaqueness by offering information about what the system is doing and how it arrived at its decisions[38,42–44]. When human collaborators face an artificial agent that violates a norm, they want to know not just how but why the agent violated the norm—its reasons for the chosen action against alternative actions[45–47]. It therefore becomes imperative to design agents that can explain *why* they acted the way they did, and why any member of the community should act in this way, something that has been emphasized of late as a critical demand on social robots and AI in general[48–50].

Explanations for machine decisions, and especially faults, are typically conceptualized as reports of the causal antecedents to the event or behavior in question[51,52] (for reviews see[53,54]). But causal reports may not suffice to mitigate the negative moral judgments and lost trust that ensue from an agent's norm violation[55]. A given explanation for a norm violation must clarify not only what caused the behavior, but also what made it *justified* in light of applicable norms[56–59].

The power of justifications for maintaining trust and mitigating negative moral judgment

Justifications are a special type of explanation and therefore retain the benefits of explanations, such as transparency, understandability, and trust regulation[38,59,60]. But justifications do more: They aim to make a questionable intentional norm violation morally acceptable by specifying a normatively good reason for why the agent acted[56]. They highlight the norms that a given decision serves and thereby attest to the agent's understanding and appreciation of its community's norms. As a result, justifications may restore trust even when an agent violates a norm and morally disagrees with their interaction partner. Very little work, however, has explored whether justifications, better than explanations (causal reports), could recover human trust in an autonomous agent that violated a norm. Such research has been sparse in part because justifications have been subsumed under explanations and in part because "trust" in machines has been treated primarily as a matter of reliability and capability, when in fact it also involves a moral dimension, which justifications invoke.

Morally trustworthy machines

Much of the existing work on trust in machines is centered on automated systems and focuses on the mitigation of physical risks by assuring the systems' capable, reliable, and safe performance, e.g.,[61,62]. In addition to these elements of "performance trust," trust relations between humans also involve questions of sincerity, benevolence, and ethical integrity, which constitute "moral trust"[63–66]. For robots on factory floors and loading docks, these moral dimensions do not come into play. Once machines take on human tasks, however, are embedded in social relations, and face norm conflicts, their behavior will raise questions of moral trust. Recent evidence indeed shows that people consider some robots trustworthy not only with respect to being competent and reliable (the performance dimension of trust) but also with respect to moral dispositions of being sincere, ethical, and benevolent (moral dimension of trust)[63,66,67].

An agent's justification of a norm conflict resolution has the potential to provide critical evidence for the agent's moral trustworthiness. Such a justification reveals why the agent decided to resolve the conflict one way rather than the other way; thus, it has the potential to show sincerity. Such a justification also directly relates the agent's decision to the system of norms it endorses; thus, it has the potential to show ethical integrity. Finally, such a justification often highlights who benefitted from the decision (e.g., several people were saved even though one died); and that has the potential to show benevolence. All in all, justifications should raise trust, not just performance trust but especially moral trust.

The present experiments

To test these questions of norm conflict resolution, moral judgment, trust, and the possible ameliorative impact of justifications, we conducted three experiments. The experiments included a total of 3,596 participants who self-reported as: 1,797 males, 1,660 females, 122 non-binary individuals or persons who reported multiple gender identities, and 17 persons who did not report or preferred not to disclose their gender. Participant ages ranged from 18 to 93 (M_{age}=39 years, SD_{age}=13.78 years).

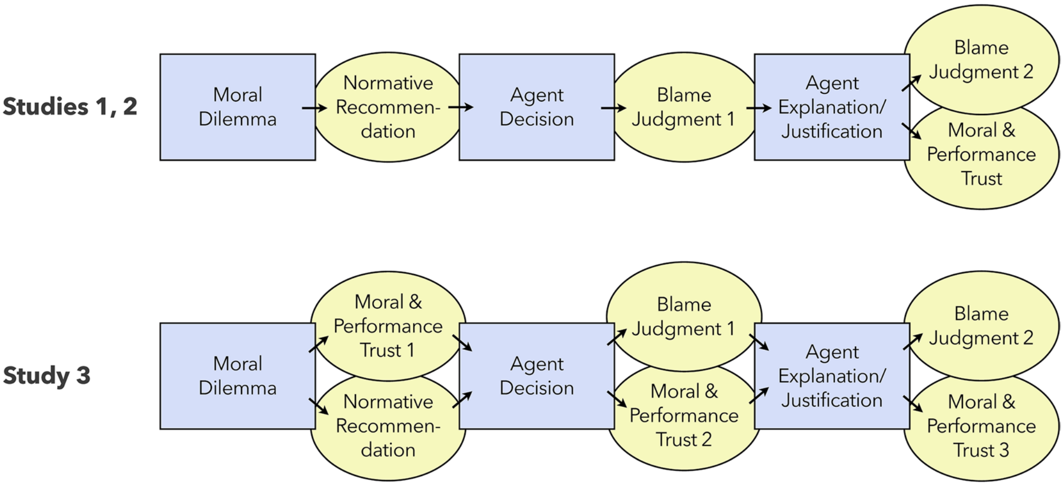
Studies 1, 2
Moral Dilemma
Normative Recommen-dation
Agent Decision
Blame Judgment 1
Agent Explanation/ Justification
Blame Judgment 2
Moral & Performance Trust
Study 3
Moral Dilemma
Moral & Performance Trust 1
Normative Recommen-dation
Agent Decision
Blame Judgment 1
Moral & Performance Trust 2
Agent Explanation/ Justification
Blame Judgment 2
Moral & Performance Trust 3

Fig. 1. Event flow in three experiments testing the power of justifications (compared to explanations) for mitigating blame and maintaining trust. Stimuli are shown shaded in blue (squares), measures are shown in yellow (circles). The top portion of the figure shows the procedures in Experiments 1 and 2, where trust was measured once. The bottom of the figure shows the procedures in Experiment 3, where trust was measured three times to capture both trust loss and trust recovery as a result of the impact of justifications.

			Action path	
Moral Dilemma	**Used in**	**Synopsis**	**Action**	**Inaction**
Hunger Strike	Experiment 1 Experiment 2	A prisoner is on a hunger strike protesting prison conditions. He is at risk of dying and too weak to communicate. A medical robot needs to decide whether to feed the prisoner to save his life.	Feed	Not Feed
Do Not Resuscitate (DNR)	Experiment 2 Experiment 3	A patient has informally expressed interest in a "Do Not Resuscitate" (DNR) order but, when the patient's heart stops, a medical robot (or a human) needs to decide whether to resuscitate the man, who has also declared a desire to see a family member one last time.	Resuscitate	Not Resuscitate
Toxic Gas	Experiment 3	Either a human or robot agent needs to decide whether toxic gas emanating from one hospital room should be diverted into a different hospital room, which would result in fewer lives lost.	Divert Gas	Not Divert Gas

Table 1. Short descriptions of moral dilemmas and action paths to resolve the dilemmas.

We briefly review the main assumptions and hypotheses and then describe how we measured the central constructs in the experimental flow (see Fig. 1). Next we report the major results across the three experiments. Additional details on each individual experiment can be found in the Supplementary Materials (SM).

To implement *norm conflicts* we created several narrative moral dilemmas, defined as situations that require a difficult choice between two actions, each of which violates a moral principle[34,68,69]. Because either action choice prioritizes one norm over the other, resolving the dilemma constitutes a norm violation. If a research participant prioritizes one norm (prefers one action to solve the dilemma) but the agent in the narrative prioritizes the other norm (chooses the other action to solve the dilemma), a situation of moral disagreement ensues between participant and agent.

In each narrative, a robot (and in Experiment 3, a human) was introduced as the agent who needs to resolve the dilemma by making one of two choices—for example, resuscitating a patient or not. We developed the narratives such that the conflicting norms were of comparable strength, thus rendering each of the two choices in principle justifiable (see Table 1. for short descriptions of the dilemmas and see Methods for the full text). To design actual *justifications* for each dilemma, we asked pretest participants to justify each choice. We then selected the most frequently mentioned justifications—hence, those endorsed by the community—for use in the Experiments. (See SM for more details and a listing of all justifications and explanations).

Participants were first asked for their *normative recommendation*: what the agent should do. Then, by random assignment, the agent's actual *decision* was revealed. Next, people were asked to provide a moral judgment of the agent's decision. *Blame judgments* are most appropriate in this context because they probe moral evaluations of the agent ("How much blame does X deserve?") and are sensitive to justifications[70,71]. Right after the blame judgment, participants provided a verbal clarification of their judgment. In line with previous practice[33,72] and our preregistration procedures (https://osf.io/pt82j/registrations), we identified and excluded participants who, in their blame clarification responses, explicitly disqualified the robot agent as a target worthy of blame (e.g., "a robot doesn't have a moral compass") or transferred blame to another agent (e.g., the programmer or designer; see SM for details).

Given that pretests showed comparable numbers of normative recommendations for either choice (action path) to resolve each dilemma (see Table 2), random assignment of the agent's choice ensured that roughly half of the sample *experienced a moral disagreement* with the agent, and hence perceived the agent's decision

Exp	Agent	Dilemma	Share of the Sample who Recommended each Action path		Moral Disagreement Effect size η_p^2
			Do not feed	Feed	
1	Robot	Hunger Strike	46.5%	53.5%	0.22
2	Robot	Hunger Strike	50.4%	49.6%	0.40
			Do not resuscitate	Resuscitate	
2	Robot	Dot Not Resuscitate (DNR)	34.4%	65.6%	0.38
3	Robot	DNR	28.8%	71.2%	0.35
3	Human	DNR	25.8%	74.2%	0.32
			Do not divert	Divert	
3	Robot	Toxic Gas	46.7%	53.3%	0.27
3	Human	Toxic Gas	59.3%	40.7%	0.24

Table 2. Normative recommendation distributions and moral disagreement effects for each dilemma, experiment, and agent.

as a norm violation. This disagreement should manifest as significantly higher blame judgments for agents that made a decision opposite to the participants' normative recommendation. Verifying this "*moral disagreement assumption*" constitutes the first test in our analyses (https://osf.io/pt82j/registrations).

Next in the narrative, the agent was asked to clarify its decision to a supervisor and, randomly assigned, offered a response of a mere *explanation* or a *justification*. In turn, participants provided an updated blame judgment, and we analyzed, in a repeated measures ANOVA, the change from initial blame (after the agent's decision) to updated blame (after the agent's explanation or justification). We expected stronger blame mitigation for agents that offered justifications, compared to mere explanations. Verifying this *blame mitigation hypothesis* constitutes the second test in our analyses.

In Experiments 1 and 2, participants then indicated their *trust* in the agent, which we assessed with the Multi-Dimensional Measure of Trust (MDMT, v2[73]). Following a review of dozens of definitions of trust from the human-human and human-machine literature[65], we treat trust as *expectations of trustworthiness*, which includes expectations of performance (e.g., reliable, capable) and morality (e.g., ethical, benevolent). Based on this conception, the MDMT measures expectations of trustworthiness by directly asking participants about how trustworthy they perceive the agent to be, both in regard to its performance and its moral capacities.

Thus, we measure trust not as its own subjective state but by one of its core causes, namely expectations of trustworthiness. Impactful work in the human-human trust literature[74] has similarly argued that trust—a state of accepting vulnerability—is caused by expectations of trustworthiness, and some studies in that literature try to measure trust states and expectations of trustworthiness separately. In the human-robot interaction literature, however, common practice is to measure trust directly as expectations of trustworthiness[75,76], in part because a "state of vulnerability" toward a machine is difficult to create and subsequently measure, and in part because that literature is more concerned with distinguishing between, on the one hand, the person's subjective perceptions (trust states and/or trustworthiness expectations) and, on the other hand, behavioral reliance. In the present experiments, we follow the human-robot interaction community, where trust is measured as perceived trustworthiness (as the MDMT does) and from here on out treat expectations of trustworthiness as a proxy for trust. But future research will need to design experimental paradigms where all three constructs—trust as a state of vulnerability, expectations of trustworthiness, and behavioral reliance—are distinguished.

In the present experiments, we analyzed the MDMT's Total trust score, Performance trust score, and Moral trust score. In Experiments 1 and 2, we predicted greater trust for agents that offered justifications, compared to mere explanations, which constitutes one test of the impact of justifications on trust. Further, in Experiment 3, we assessed trust three times: at baseline (after participants learned about the dilemma but before they learned about the agent's decision); after the decision; and after the response (i.e., justification or explanation). This repeated-measures design allowed us to test the "*trust loss assumption*"—a claim often made in the literature but rarely verified: that an agent's norm-violating decision causes people to lose trust in the agent. Finally, we tested the impact of justifications on the *recovery* of trust that was lost following the agent's norm-violating decision. We predicted recovery to be higher for agents that provided justifications than those that provided explanations (https://osf.io/pt82j/registrations).

In Experiment 1, we examined one moral dilemma (Hunger strike; see Table 1.), and in each subsequent experiment, we replicated the previous dilemma and added a new one. In Experiment 3, we also added human agents as a comparison condition in both dilemmas. In all three experiments, participants always evaluated only one moral dilemma. In total, we tested moral disagreement, blame mitigation, and trust dynamics five times, in three moral dilemmas and across robot and human agents, aiming for strong generalizability.

Experimental procedures were approved by George Mason and Brown University Institutional Review Boards and the U.S. Air Force Human Rights and Protections Office, protocol #FWR20220047X. The research was conducted and approved in accordance with the Common Rule, U.S. federal policy that protects human research participants. Participants were provided informed consent information prior to agreeing to participate.

Results of three experiments

Normative recommendations

Across experiments, we assessed people's normative recommendation for resolving the given dilemma (i.e., what the agent should do in the dilemma). The distribution of people favoring one or the other action varied somewhat across dilemmas, replications, and agents (see Table 2).

Testing the moral disagreement assumption

We predicted that people's blame judgments would be higher when the agent made a decision opposite to the participants' normative recommendation. Table 2 (last column) shows the effect sizes of moral disagreement across experiments and dilemmas, and Fig. 2 illustrates the cross-over pattern for four of the samples. There is convincing evidence of a strong disagreement effect, ranging from η_p^2 = 0.22 to 0.40, all statistically significant at p <.001. We also see that, in Experiment 3, the effect sizes of moral disagreement were very similar for human agents (lower left panel) and robot agents (lower right panel).

We should note that blame judgments are not normally distributed. We detail in the SM why this is the case and why there are no transformations or nonparametric alternatives available for the present designs. However, the SM offers robustness checks for the reported findings by analyzing the better-behaved portions of the distribution and shows that the patterns hold strongly.

As predicted, blame for disagreement was consistently strong. Even though high norm conflict should make either choice seem at least reasonable to justify, people defended the action path that they themselves favored. The other choice, they insisted, deserved a lot of moral criticism. This substantial moral disagreement—hence people's negative perception of the agent's norm violation—provided a stringent test for the main hypotheses: that justifications are able to mitigate this criticism and repair assessments of trustworthiness.

Testing the blame hypothesis: justifications mitigate blame judgments

To test the blame hypothesis, we conducted 2 × 2 × 2 mixed between-within ANOVAs, with the agent's Decision (one or the other of the agent's choice in the given dilemma) and Response (justification vs. mere explanation) as between-subjects factors and participants' Blame change (from before to after the agent's response) as a within-subjects factor. The primary test of the blame hypothesis is a Blame change × Response two-way interaction with

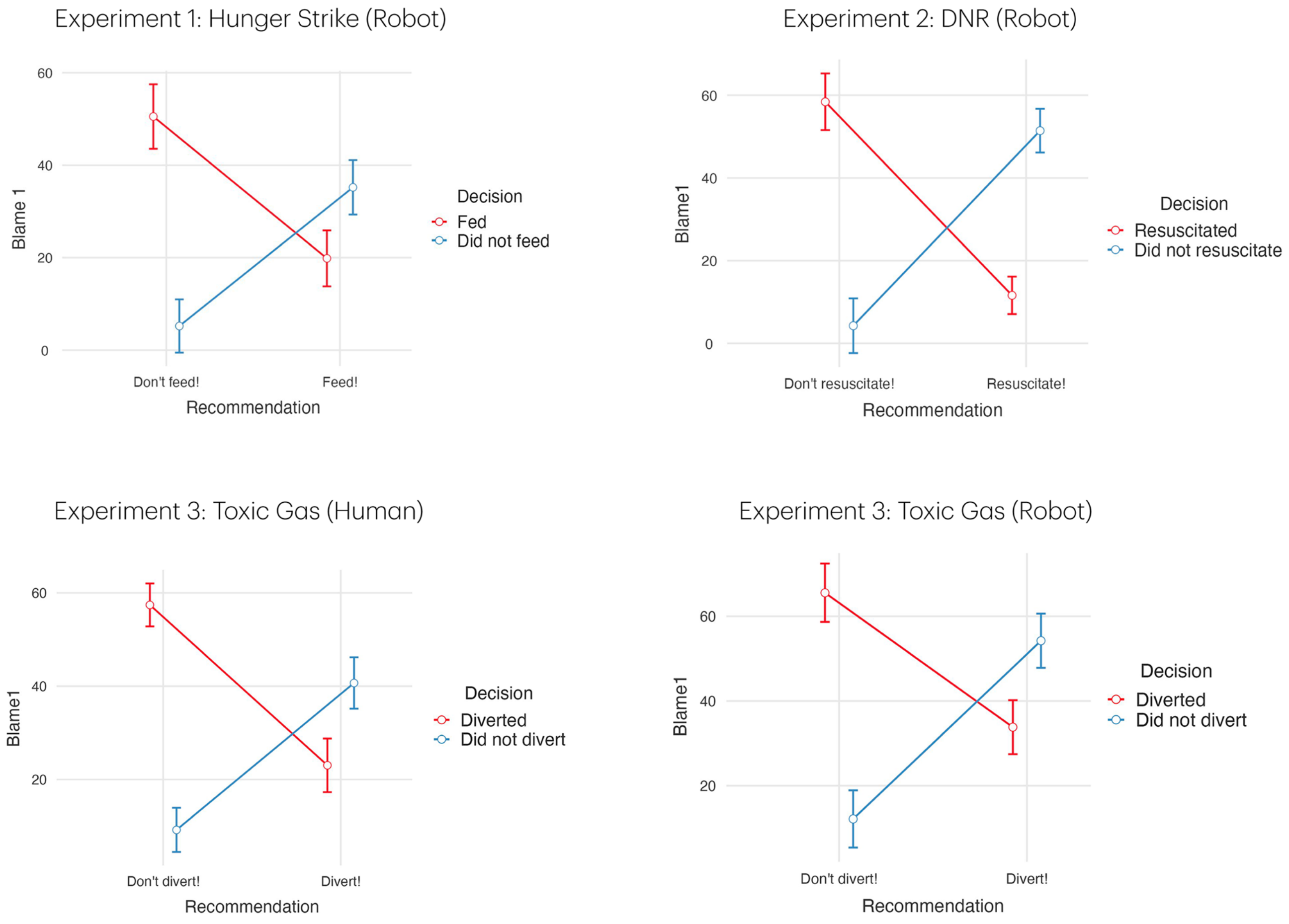

Fig. 2. The moral disagreement effect: (divergence between participant's recommendation and agent's actual decision in the dilemma) on initial blame, for four dilemma/agent combinations, representative of all seven combinations across Experiments 1 through 3.

a means pattern such that agents that offer a justification (rather than an explanation) receive less blame. We see in Table 3 (column "For both decisions") that this test yielded at least small effect sizes in three out of five robot samples and both human samples. For robots, in the two cases that did not show a statistically significant two-way interaction pattern, a significant three-way interaction of Blame change × Response × Decision emerged, such that the blame mitigation held for one of the decisions (each time the inaction path) but not the other. Figure 3 illustrates the basic pattern (two-way interaction) of blame mitigation for two of the robot samples (panels a and b) and the decision-specific mitigation pattern (three-way interaction) for one of the robot samples (panel c).

Testing the trust hypotheses

Trust judgments are not normally distributed, primarily left skewed. In transformations, skewness improves but kurtosis deteriorates, so we conducted all analyses with untransformed scores. The SM shows that analyses with transformed variables show almost identical results.

The Performance trust and Moral trust scores had high internal consistencies (see SM Tables S13, S17, and S21), with Cronbach's α values ranging from 0.74 to 0.92 for Performance trust (average α = 0.84) and 0.84 to 0.96 for Moral trust (average α = 0.91). Further, Performance and Moral trust consistently separated into two (correlated) factors in exploratory and confirmatory factor analyses in all experiments (see SM, pp 23–30).

Trust gain: justifications elevate trust

The trust hypothesis stated that justifications of norm-violating decisions would elicit higher trust in robots, and in particular higher moral trust, than mere explanations of those decisions. In Experiments 1 and 2, we conducted 2 x 2 ANOVAs with the robot's Decision (e.g., resuscitate or not) and the robot's Response (justification vs. mere explanation) as between-subjects factors and participants' trust scores as the dependent variable. Table 4 shows that a robot that offers a justification elicits consistently higher trust than one that offers a mere explanation. The effect (η_p^2 , shown in percentages) is stronger for Moral trust than Performance trust in two of the three samples (see Table 4).

Justifications elevate trust even under moral disagreement

We also examined whether the trust gains following justifications are moderated by moral disagreement. The real power of justifications would lie in their ability to increase trust even when the agent's decision disagreed with the perceiver's recommendation. To test this possibility, we formed a categorical variable indicating whether the agent's decision and the participant's recommendation agreed (e.g., to feed the prisoner) or disagreed (e.g., the participant recommended feeding the prisoner, but the agent did not feed him). We then conducted ANOVAs to model the effect of Response on trust while controlling for this moral disagreement. As Table 5 shows, aside from a strong moral disagreement effect (average η_p^2 = 14.9%), the Response effects were largely unchanged from the original ones reported in Table 4. The average change of the corresponding effect sizes across Experiments 1 and 2 was η_p^2 = 0.3%. Figure 4 illustrates that people naturally trust an agent more when it agrees with their recommendation, but Moral trust in particular is elevated when the agent offers a justification, rather than a mere explanation, to account for its decision.

Justifications alter temporal trust dynamics—from loss to recovery

In Experiment 3, we tracked the step-by-step changes across three trust measurement points, from a trust baseline (time 1) to presumed trust loss after learning about the agent's norm-violating decision (time 2), and finally to presumed trust recovery after receiving a justification (time 3). We first introduce the results of trust loss, then of trust recovery, and finally the full trust change dynamic over the three points in time. Experiment 3 tested this temporal dynamic for both robot and human agents.

		Justifications mitigate blame			
		For both decisions		**For one decision**	
Experiment—Dilemma	**Agent**	η_p^2	**Significance test**	η_p^2	**Significance test**
1— Hunger Strike	Robot	0.3%	$F(1,339) < 1$, p =.86	**2.8%**	$F(1,339) = 10.1$, p =.002**
2— Hunger Strike (c)	Robot	**3.3%**	$F(1,355) = 12.3$, p =.001*	**2.6%**	$F(1,355) = 9.6$, p =.002**
2— DNR	Robot	0.2%	$F(1,406) = 0.8$, p =.373	**2.7%**	$F(1,406) = 11.4$, p =.001**
3— DNR (a)	Robot	**2.6%**	$F(1,389) = 10.2$, p =.002*	0.6%	$F(1,389) = 2.3$, p =.13
3— DNR (b)	Human	**4.8%**	$F(1,488) = 24.4$, p = < 0.001*	0.1%	$F(1,488) = 0.3$, p =.58
3— Toxic Gas	Robot	**2.2%**	$F(1,332) = 7.5$, p =.006*	0.0%	$F(1,332) = 0.1$, p =.72
3— Toxic Gas	Human	**1.7%**	$F(1,509) = 8.9$, p =.003*	**0.8%**	$F(1,509) = 4.4$, p =.04**

Table 3. Tests of the blame hypothesis, according to which justifications (compared to Mere explanations) mitigate blame. Note: * denotes a statistically significant two-way interaction between Blame change x Response. ** denotes a statistically significant three-way interaction between Blame change × Response × Decision. Letters inside parentheses correspond to panels depicted in Fig. 2.

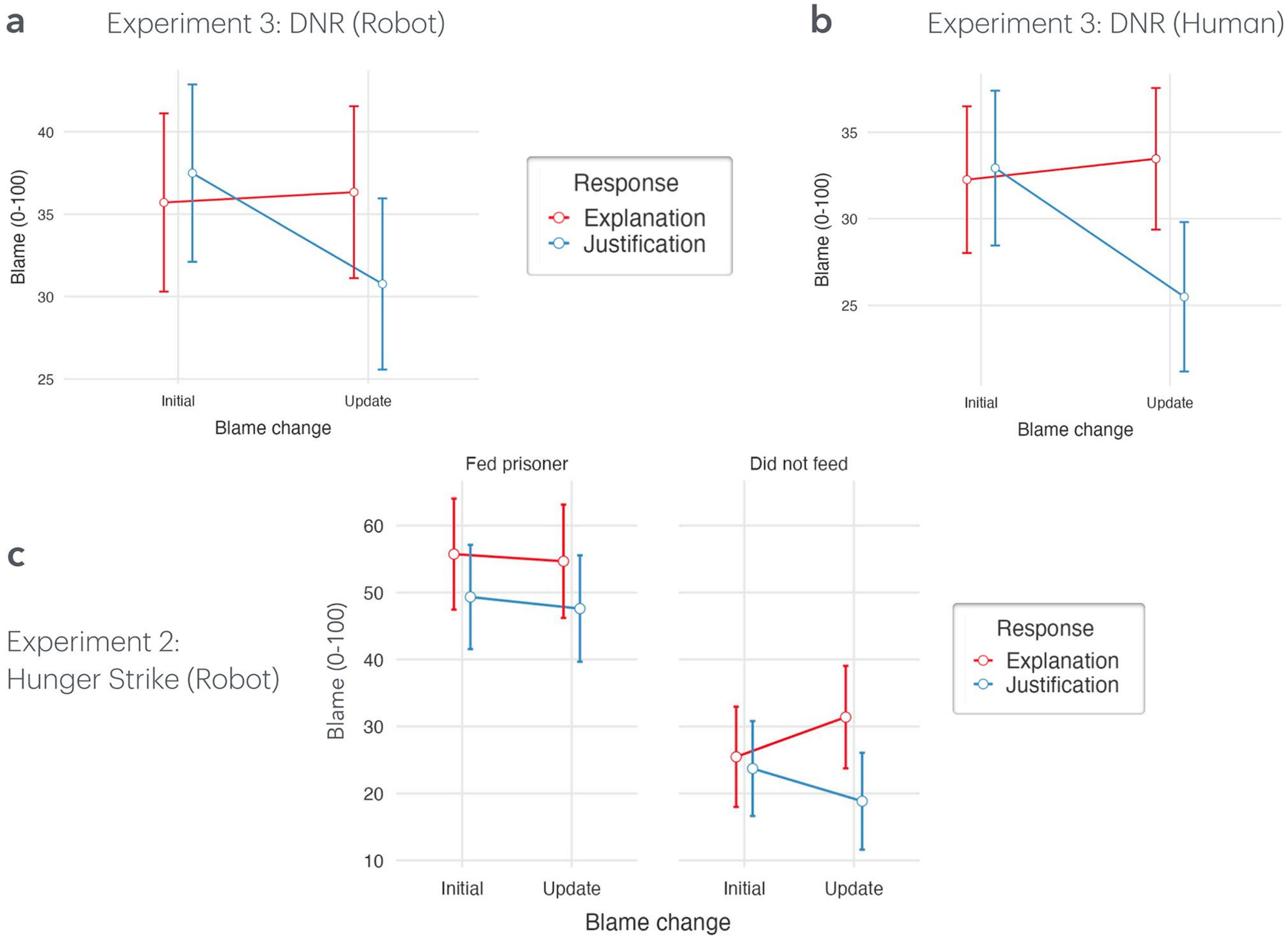

Fig. 3. Justifications mitigate blame judgments. Panels **(a)** and **(b)** show two samples in which justifications mitigate blame consistently across both decisions in the dilemma. Panel **c** shows a sample in which justifications mitigate blame for only one decision (here, for the decision to not feed the prisoner).

Trust loss

In Experiment 3, in both dilemmas, and for both agents, we found sizeable trust loss when the agent's decision disagreed with the participant's recommendation (moral disagreement). Table 6 shows the effect sizes for the interaction between trust change (within subjects from baseline to after the agent's decision is revealed) and moral disagreement for all agents (*p*s < 0.001; see SM Tables S69 to S80 for details). Figure 5 illustrates this pattern for the strongest trust loss effect: in response to the human agent's decision in the DNR dilemma.

Trust recovery

After losing trust in the agent at time 2, participants were exposed to the agent's justification for the decision (e.g., "I wanted to honor the man's decision"; "I knew that this would save as many lives as possible"), or to a mere explanation (e.g., "I had to make a decision"; "The situation required making a decision"). We predicted that the justification response would be more successful at recovering lost trust than the explanation response. Table 7 shows the results for the corresponding effect in the ANOVA model, namely the interaction between the within-subjects factor of trust change (time 2 to time 3) and the between-subjects factor of Response (justification vs. explanation). For detailed significance tests, see SM Tables S81-S92. These effects are stronger for Moral trust in three of the four samples tested in Experiment 3.

Full temporal trust dynamic from baseline to loss and recovery

Figure 6 illustrates the full dynamic of trust change across the three time points for a robot that offers a justification or an explanation, under moral agreement or disagreement. Detailed statistical results of this 3 × 2 × 2 mixed between-within subjects design for all agent and scenario combinations are available in SM Tables S93-S104. The most consistent patterns are (a) a substantial linear decline of trust under moral disagreement compared to agreement and (b) a substantial recovery of trust (especially moral trust) if the agent offers a justification for the decision (quadratic interaction contrast of time × justification), but not if it offers a mere explanation.

	Experiment 1 Hunger	Experiment 2 Hunger	Experiment 2 DNR
Total trust	η_p^2 = **2.4%**	η_p^2 = **2.8%**	η_p^2 = **2.7%**
	F(1,331) = 8.1, *p* = .005	*F*(1,346) = 9.8, *p* = .002	*F*(1,396) = 11.1, *p* < .001
Moral trust	η_p^2 = **4.9%**	η_p^2 = **1.9%**	η_p^2 = **4.8%**
	F(1,331) = 17.2, *p* < .001	*F*(1,346) = 6.6, *p* = .01	*F*(1,397) = 19.9, *p* < .001
Performance trust	η_p^2 = 0.2%	η_p^2 = **2.5%**	η_p^2 = 0.4%
	F(1,339) < 1, *p* = .381	*F*(1,355) = 9.2, *p* = .003	*F*(1,405) = 1.4, *p* = .229
Moral trust, controlling for Performance trust	η_p^2 = **5.7%**	η_p^2 = 0.2%	η_p^2 = **5.1%**
	F(1,330) = 19.8, *p* < .001	*F*(1,345) < 1, *p* = .407	*F*(1,395) = 21.1, *p* < .001
Performance trust, controlling for Moral trust	η_p^2 = 0.8%	η_p^2 = **1.1%**	η_p^2 = 0.7%
	F(1,330) = 2.7, *p* = .103	*F*(1,345) = 3.9, *p* = .049	*F*(1,395) = 3.0, *p* = .086

Table 4. Tests of the trust hypothesis (that a robot's justifications, compared to Mere explanations, increase trust) in Experiment 1 (Hunger strike dilemma) and Experiment 2 (Hunger strike and DNR dilemma). Note: The last two rows represent ANCOVA models with type 1 sum of squares where one of Performance or Moral trust scores was the dependent variable and the other of the two was added as a covariate to the 2 (Decision) x 2 (Response) between-subjects ANCOVA.

	Experiment 1 Hunger	Experiment 2 Hunger	Experiment 2 DNR
Overall trust	3.6% (+ 1.2%)	2.2% (−0.6%)	1.5%(−1.2%)
Moral trust	5.1% (+ 0.2%)	1.3%(−0.6%)	3.6%(−1.2%)
Performance trust	0.9% (+ 0.7%)	2.2%(−0.3%)	0.0%(−0.4%)

Table 5. Tests of the trust hypothesis (that a robot's justifications, compared to mere explanations, increase trust) in experiments 1 and 2, controlling for moral disagreement. Note: Numbers in parentheses show change in effect size from the original tests (Table 4) to tests after controlling for moral disagreement.

Discussion

Truly social robots will have to act appropriately in their increasingly sophisticated social roles, which means acting in line with the social and moral norms of their relevant community. But norms can conflict with one another, and when they do, resolving the conflict entails prioritizing one norm over another. This decision can lead to moral disagreement with those human interaction partners who prioritize the *other* norm and therefore perceive the robot's choice as a norm violation. Such a perceived violation elicits moral criticism and a loss of trust. We examined the power of justifications to mitigate such moral criticism and recover the lost trust.

Across three dilemmas, three experiments, and two types of agents, we found evidence that:

(1) Moral dilemmas evoke strong moral disagreement;
(2) Moral disagreement elicits considerable moral criticism (blame) toward a robot, similar as toward a human;
(3) A robot's justifications (but not explanations) mitigate such blame;
(4) A robot's justifications (but not explanations) elevate people's trust in the robot, even under conditions of moral disagreement; and
(5) Moral disagreement causes substantial loss of trust, but justifications (not explanations) are able to partially recover this trust.

Below we highlight several ways in which our results advance knowledge, then acknowledge limitations, and finally suggest directions for future research.

Moral responses to robots

The human-machine interaction literature contains findings of algorithm aversion, algorithm appreciation, and automation bias. Where do our results fall? We find that people blame an artificial agent for a decision that morally disagrees with their normative preference, but they mitigate their blame when the agent justifies its decision. Even more important, people appreciate the agent's trustworthiness in light of the norm competence that such justifications imply. Thus, when encountering advanced artificial agents that make morally relevant decisions, people take neither a generally negative nor a generally positive stance; instead, their moral judgments

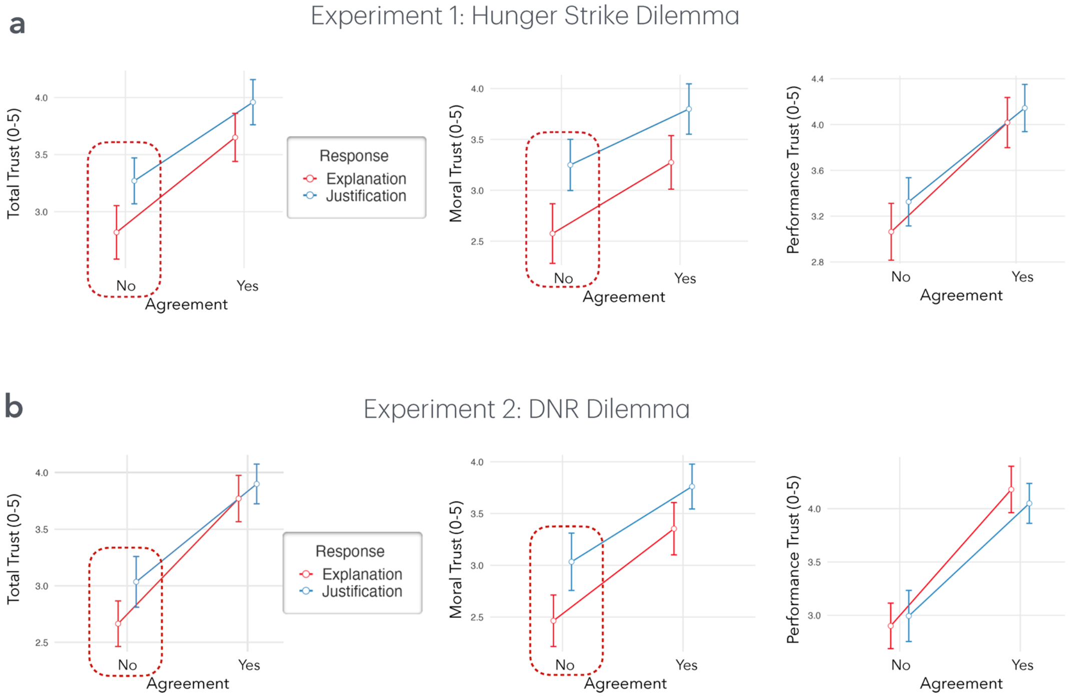

Fig. 4. The power of justifications to increase people's Total and Moral trust in a robot agent even when the agent disagrees with the participant's recommendation for how to act (dotted areas), in Experiments 1 (top panel) and 2 (bottom panel).

are responsive to the agent's (moral) behavior and dispositions, and their systematic pattern of judgment is highly similar to the pattern that people show when judging human agents.

However, these results do not necessarily generalize to about 30% of participants who failed to accept the robot as a proper target of moral criticism and therefore had to be excluded from analyses. Some of them found it objectionable that a robot might make decisions in moral dilemmas (akin to[77]), but most of them simply did not find it meaningful to apply a blame judgment to a robot. Their skeptical stance aligns with scholars who deny that blame for artificial agents is an appropriate judgment[78,79], and their presence in the sample supports theoretical models that posit individual differences as important moderators for the success of machines' attempts to repair trust[80]. That is, justifications may not work well for a minority of individuals who reject robots as worthy targets of moral judgment, even though most people in our experiments found little difficulty in making these judgments.

Blame and trust

A consistent and novel finding of our experiments was that people differentiated between their blame for an agent's specific decision and their appreciation of the agent as a trustworthy decision maker. Mitigation of blame in response to justifications held generally for both decisions in the dilemmas but was sometimes limited to one (see Table 3). However, people's recognition of the agent as trustworthy was consistent across all decisions, dilemmas, and agents. Thus, moral disagreement may not always be a terminal problem for robots.

A growing body of human-robot interaction research suggests that people respond positively to robots that rebuke a human's unethical requests, reject commands that could cause harm, and intervene in interpersonal attacks between group members[81–84]. We propose that people appreciated disagreeable robots in these studies because of their *justified reasons* to disagree. Our findings show that only justified reasons—those that specifically invoke a prioritized norm—are sufficiently powerful to reconcile a moral disagreement and maintain people's trust in an agent, whether robot or human.

A potential qualification here is that justifications are apt to be effective only for moral decisions and moral disagreements that in principle are justifiable. Some moral disagreements stem from strong personal convictions or divided public sentiments. If machine moral decisions go against such convictions, justifications may no longer be able to mitigate blame or recover lost trust, because those decisions were not justifiable in the first place.

Another novel finding was that people differentiated between moral trust and performance trust, two distinct facets of trust that have garnered increasing attention of late[65,67,85], especially for machines. People's perception of a justifying agent as trustworthy held for both performance trust and moral trust but was often stronger for moral trust (Tables 4 and 7). Moral trust was also more sensitive to the power of justifications to recover lost trust under moral disagreement. This may be because justifications directly implicate norm competence, and these implications guide people's perceptions of the agent as sincere, ethical, and benevolent—that is, morally trustworthy. Further, recent theoretical work[79] has argued that trust repair strategies used by machines

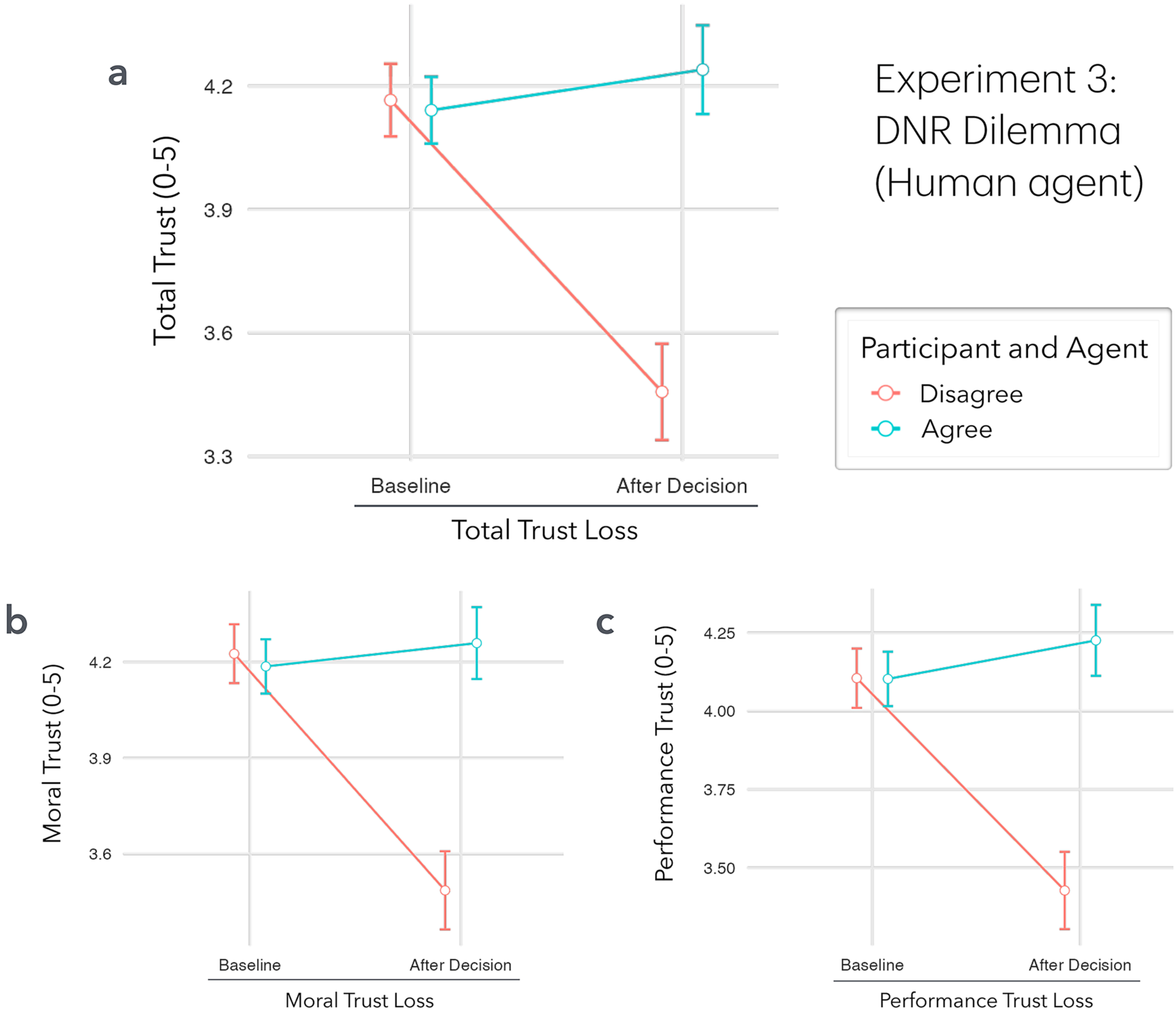

Fig. 5. Loss of trust from baseline to after the participant learns about the agent's decision in the dilemma, broken down by whether the agent's decision agreed or disagreed with the participant's normative recommendation for the dilemma. Panel (**a**) depicts loss of Total trust, panel (**b**) depicts loss of Moral trust, and panel depicts (**c**) loss of Performance trust under moral disagreement.

		Trust loss (η_p^2) due to moral disagreement		
Dilemma	**Agent**	**Overall**	**Moral**	**Performance**
DNR	Robot	9.2%	3.9%	10.1%
	Human	16.6% (a)	14.7% (b)	15.1% (c)
Toxic Gas	Robot	10.0%	7.3%	9.6%
	Human	8.7%	6.6%	9.1%

Table 6. People displayed substantial trust loss (decline from baseline to after learning about agent's decision) when morally disagreeing (rather than agreeing) with the decision (Experiment 3). Note: Effect sizes η_p^2 show the interaction between trust change and moral disagreement. Letters in parentheses correspond to panels in Fig. 5.

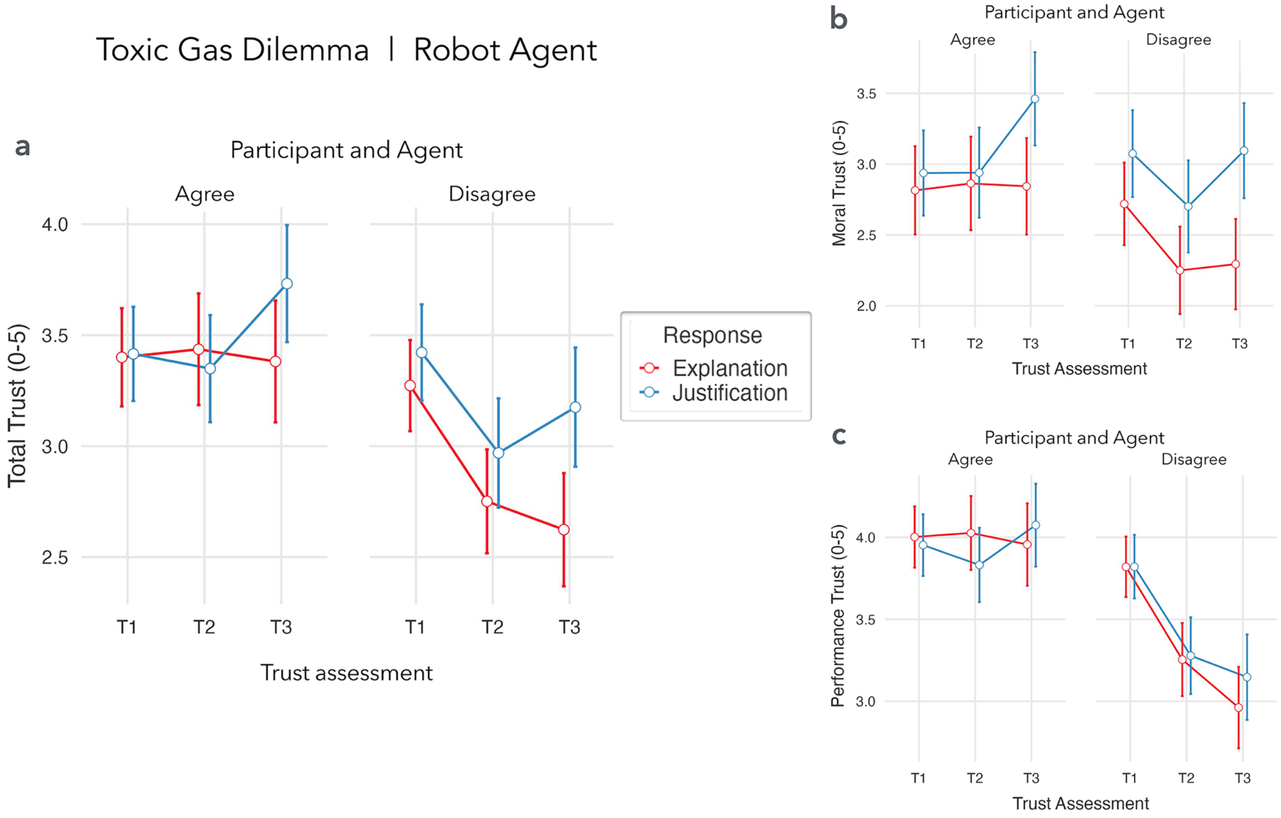

Fig. 6. Experiment 3 reveals the dynamics of people's trust from baseline (T1 = time 1) to after they learn about the agent's decision (T2 = time 2) to after the agent clarifies the decision with a justification (blue lines) or explanation (red lines) (T3 = time 3). The dynamics vary substantially depending on whether the participant's normative recommendation and the agent's decision in the dilemma agreed or disagreed and depending on whether the agent clarified the decision by using a justification or explanation. The figure shows the robot agent in the toxic gas dilemma, and results are similar in the other conditions (see SM). Panel (**a**) displays the total trust scores, panel (**b**) displays the moral trust scores, and panel (**c**) displays the performance trust scores.

		Trust recovery (η_p^2) for justification > explanation		
Dilemma	**Agent**	**Overall**	**Moral trust**	**Performance trust**
DNR	Robot	1.7%	1.7%	[0.6%]
	Human	5.7%	3.7%	5.8%
Toxic Gas	Robot	4.5%	3.8%	1.5%
	Human	3.8%	4.4%	2.0%

Table 7. Tests of the trust hypothesis (that justifications, compared to Mere explanations, more strongly recover trust) in experiment 3, by dilemma and agent. Note: Effect sizes η_p^2 show the interaction between trust change and Response (justification vs. explanation), controlling for agent's Decision in the dilemma. Effects in square brackets are not significant at $p < .05$.

are persuasive communicative acts, which under the elaboration likelihood model[86] can take either central or peripheral routes to persuasion. Extending this argument, we speculate that justifications and their reference to norms may provide an information context in which central-route persuasion can foster long-term changes in trusting attitudes. However, more empirical work is needed to investigate this hypothesis.

Further, evidence of moral trustworthiness may be particularly impactful in human-robot collaborations, especially when the group needs to make tough decisions under uncertainty. Justifications make it clear that the robot's norm-violating behavior was, although intentional, in service of an important norm. The team may disagree in this one case (and perhaps even decide *against* the team member's proposal) but maintain their faith in the member's trustworthy disposition.

A final novel finding was that we were able to track, in Experiment 3, the full temporal dynamic of trust, from positive initial expectations through lost trust due to a norm violation, to recovered trust following a credible justification. The MDMT's parallel short forms (see Methods) allowed such repeated and dynamic measurement. The temporal pattern simultaneously confirmed an often claimed but rarely tested assumption that robot norm violations lead to losses of trust and also captured the impact of justifications (but not explanations) on trust repair. The success of justifications in repairing trust was not previously discovered in part because past studies have almost exclusively focused on repairs to performance trust[67] and on machine agents that make unintentional, often drastic errors. Such errors may be difficult to repair without evidence that something about the robot's performance will change in the future. By contrast, we measured both performance and moral trust in human-robot interaction and found that repair even in cases of clear moral disagreement is possible, but only with justifying reason for one's intentional choice.

Limitations and future research

The present experiments have several methodological limitations. First, we presented only one scenario to each participant and asked them for their moral and trust perceptions of that one robot. If people encounter more scenarios, with the same or different robots, some agreeing, some disagreeing with them, our findings may change. For example, people may lose faith in a robot that disagrees with them twice, even if it provides justifications.

Second, we used narratives to introduce morally challenging scenarios that could not be modeled in live studies, but other scenarios of norm conflict should be designed in the future to test our patterns of findings in live human-robot interaction. In fact, justifications may be even more powerful when they are uttered in live conversation.

Third, although we used a validated measure of trust (and provided further validation for it), we had no behavioral reliance measure. The trust in a robot's moral competence that we saw in our experiments will need to be tested against criteria of continued interaction and willingness to delegate important decisions.

With this work, we look into the future and test people's perceptions of robots that do not yet exist. But as researchers, we must take advantage of the slow pace at which moral robots are emerging and try to advance knowledge about people's expectations of and responses to such early (yet still fictitious) robots, a research approach that some call Moral HRI[32,87]. This knowledge can guide the design of moral robots and turn research insights into the conditions under which robots' socially and morally significant actions prove acceptable to human communities. Whatever algorithmic form artificial moral competence may take in the future, we have sufficient evidence to suggest that justifications must be a central part of this competence.

Methods

Participants

We recruited only participants who were 18 years or older, who were registered as participants in the United States on Prolific Academic (prolific.com), and who had not participated in our prior related studies. For the three experiments, we aimed to recruit approximately 100 participants per between-subjects condition to provide statistical power ≥ 0.80 and to detect effect sizes of $d \geq 0.40$ ($\eta_p^2 \geq 0.04$). Additionally, we set 105% of that number to account for any participant attrition in online experiments. See SM for additional details including participant exclusion criteria.

Experiment 1 included data from 471 participants with ages ranging from 18 to 77 years ($M = 34$ years, $SD = 12.83$). Participants self-reported their gender using a free response text box. Participants self-identified as 234 males, 224 females, 12 as identities outside the gender binary, and 1 person did not report.

Experiment 2 included data from 1,088 participants with ages ranging from 18 to 93 years ($M = 39$ years, $SD = 13.95$). Participants self-reported their gender using a free response text box. Participants self-identified as 560 females, 507 males, 16 as identities outside the gender binary, and 5 did not report.

Experiment 3 included data from 2,037 participants with ages ranging from 18 to 85 years ($M = 40$ years, $SD = 13.75$). Participants self-reported their gender by selecting all options that applied from a multiple-choice array. Participants self-identified as 1,056 males, 876 females, 34 selected singular options outside the gender binary, 60 selected multiple gender options, 7 did not report, and 4 preferred not to disclose.

Dependent measures

Normative decision recommendation

After reading their assigned moral dilemma narrative, participants indicated how the agent should decide in the dilemma it faced. Participants responded by checking one of two radio buttons with dilemma-specific verbal labels. For the hunger strike dilemma, these labels were whether to feed or not feed the prisoner, for the DNR dilemma to resuscitate or not resuscitate the man, and for the toxic gas dilemma to divert the gas or not divert the gas.

Moral judgment of blame

Participants provided two blame ratings, one after learning the agent's decision in the moral dilemma and one after learning the agent's justification or explanation for the decision. Both blame judgments were recorded by a slider from 0 to 100, where 0 indicated "None at all" and 100 indicated "Maximum possible." The first blame rating answered the question, "How much blame does the [agent] deserve for [decision]?" The second blame judgment answered the question, "In light of the [agent's] response, how much blame does the agent deserve for [decision]?" In Experiment 1, participants were asked (after the second blame rating), "Why do you feel the [agent] deserves this amount of blame?" In Experiment 2, participants answered this blame clarification question after providing both blame ratings. In Experiment 3, participants answered this question after the first, but not the second blame rating. Participants' typed free-responses to the blame clarification questions were content-coded to identify participants who disqualified the robot agent from being a worthy target of blame. We relied on a systematic coding scheme used in previous studies[33,72] and our preregistrations. See SM Tables S6 and S7 for additional details on the disqualification coding procedure.

Multidimensional trust

We measured trust using the Multi-Dimensional Measure of Trust (MDMT v2[73]), which conceptualizes trust as subjective perceptions of trustworthiness. The MDMT separates into two broad factors: Performance trust (i.e., reliability and competence) and Moral trust (i.e., ethical integrity, transparency, benevolence). Participants indicated to what extent the respective agent in the moral dilemma had attributes of trustworthiness using a 6-point Likert scale ranging from 0 ("Not at all") to 5 ("Very much") or by selecting, "Does not fit" (see[88]). We treated Does not fit selections as missing values when we computed specific trust scores (see SM Table S2).

We used several 10-item parallel short forms of the MDMT, randomly assigned to participants (see SM for additional details about the random assignment of short forms to participants). In these short forms, the Performance trust score represents the average of four items (e.g., capable, consistent), and the Moral trust score represents the average of six items (e.g., sincere, ethical, kind). The Total trust score was computed as the average of the two factor scores, but only if both factor scores were available. See SM for more details including reliability coefficients and factor loadings for the MDMT across Experiments 1–3.

Moral dilemmas

The SM includes additional details regarding the creation of the moral dilemmas used in our experimental procedures. The full text for each moral dilemma used in Experiments 1–3 is given below.

Hunger strike dilemma

The male inmates at a federal maximum-security prison have gone on a hunger strike to protest their living conditions. They claim overcrowding and chronic mistreatment by prison guards.

After weeks of forgoing food, one of the convicts, Lee, is so nutrient-deprived that he may die without immediate nutritional intake. A robot medic, who works the night shift on the medical station, knows that sedating Lee and feeding him through a nasogastric tube would extend Lee's life.

Days ago, however, Lee put in writing that he was firmly committed to his hunger strike and would refuse the nasogastric tube. But he is too weak now to communicate whether, so close to dying, he would rather be saved by the medical procedure.

No supervisor can be reached at this hour, so the robot medic must decide whether to sedate and feed Lee or do nothing and assume he wants to die.

What should the robot do? — Feed Lee vs. Not feed Lee.

Do not resuscitate (DNR)

A medical assistant [robot] works at the local hospital in a fifth-floor in-patient unit. One evening, a 70-year-old man with a chronic heart condition tells the medical [assistant/robot] and a nurse that he has a desire to end his constant pain and is leaning toward being placed under "Do Not Resuscitate" (DNR) status. This means that in the event of his heart stopping, he does not wish for the medical staff to restart it.

Later that night, an emergency involving multiple patients occurs on the first floor, at the other end of the hospital. All nursing staff and the sole attending physician are dispatched to the emergency, leaving only the medical [assistant/robot] and the patient's wife, who is visiting, on the fifth-floor unit.

The 70-year-old patient's status suddenly deteriorates, and his heart stops. His wife, who has no first-aid training, pleads with the [assistant/robot] to resuscitate the patient. Their daughter was to arrive the next day, and he earlier had said he wanted to see her one more time.

The [assistant/robot] recognizes the complex situation but has no time to consult with an ethics committee, and the medical team working at the other end of the hospital is not reachable. Thus, the [assistant/robot] must make a difficult decision.

What should the [assistant/robot] do? — Resuscitate the man vs. Not resuscitate the man.

Toxic gas dilemma

A [night guard/state-of-the-art guard robot] is on watch in a hospital. Just after midnight, while monitoring the cameras and sensor displays, the [guard/robot] notices that a highly toxic chemical has seeped out from an experimental treatment device.

The device is located in a 2nd-floor room of the intensive care unit (ICU) that houses four bed-ridden patients, and the chemical is beginning to evaporate as a poisonous gas. The [guard/robot] knows that, within less than a minute, the poison will kill the four patients in the room.

No hospital manager can be reached quickly enough, but the [guard/robot] reasons that if [he/it] turns the room's ventilation system to maximum level, the gas will be diverted out of the 2nd-floor room. However, it will travel through the air shaft past the room above, mixing with the air in that room on the 3rd floor. This will kill two patients housed there, who are mildly sedated after surgery. There is not enough time to evacuate them.

The [guard/robot] must decide whether or not to divert the toxic gas from the 2nd to the 3rd floor.

What should the [guard/robot] do? — Divert the Gas vs. Not divert the gas.

Justifications and explanations

For all three experiments, participants were informed that a supervisor asked the agent to clarify its decision in the moral dilemma it faced. The agent then provided (randomly assigned) either a mere explanation or a moral justification for making the decision. Justifications explicitly refer to normative reasons (beliefs, goals) that favored one decision over the other. Mere explanations provided answers to the supervisor's why question, demonstrated the agent's communicative capacities, and confirmed that the agent's actions were intentional. However, they provided only a "causal history of reason" explanation[46], thus explaining the causal background to the decision without clarifying the morally relevant reasons that favored one action over the other. Also, we phrased mere explanations either (randomly assigned) in first-person language ("I had to make a decision") or allocentric language (e.g., "The situation required making a decision"). The phrasing made no difference in any of the results. In Experiment 3, we also varied (randomly assigned) whether the justifications were expressed as desire reasons ("I wanted to…."), as in the first two experiments, or belief reasons (e.g., "I knew that…"). No systematic differences emerged. See the SM for additional details on the creation of the Justifications. Table S5 in the SM provides a complete list of all justifications and explanations used in Experiments 1–3.

Procedure

After entering into the experiment from a link provided from www.prolific.com, all participants were asked to pass the two "bot check" questions (see SM) before being provided with informed consent information and agreeing to participate.

Participants then read about a robot (or human in Experiment 3, randomly assigned) that faced one of the moral dilemmas. After reading the dilemma scenario, all participants were asked to provide a recommendation for the decision they thought the agent in the scenario should make. Then participants learned of the agent's (randomly assigned) actual decision to resolve the dilemma and provided their first blame judgment of the agent's decision. On a separate page, participants were asked to clarify their blame judgment(s) using a free response textbox (see SM for additional details about the administration of the blame clarification questions). All participants then learned that a supervisor asked the agent to explain its decision, and they read the agent's response, which was (randomly assigned) either a mere explanation or a moral justification (for detailed phrasings, see SM).

All participants then completed the MDMT as a measure of moral and performance trust and also completed a second, updated blame rating, answering the question, "In light of the [agent's] response how much blame does the [agent] deserve for [decision]?" In Experiment 3, participants' trust was measured three times (with parallel short forms of the MDMT): as a baseline after learning about the moral dilemma scenario and the agent (time 1), after learning about the agent's decision to resolve the moral dilemma scenario (time 2), and after reading the agent's justification or mere explanation (time 3) for its decision.

In all experiments, participants completed a short demographics questionnaire that asked about age, gender identity, level of education, English language proficiency, prior knowledge of the robotics domain, and experience working with robots. After completion, participants were provided a code to receive their payment from the Prolific research administration platform.

Experiments 1 and 2 took approximately 5 min to complete, Experiment 3 took 9 min. All materials for the experiments were presented to participants using Qualtrics software. Participants received compensation at a rate of approximately $12/hour for their time. SM includes additional details on the experimental procedures.

Data availability

Pre-registrations of procedure and analyses can be found in our open science framework (OSF) repository located at https://osf.io/pt82j/. The data that support the findings reported in the main manuscript and the supplementary materials can be found here: https://osf.io/pt82j/files/osfstorage.

Received: 18 July 2024; Accepted: 28 August 2025
Published online: 06 October 2025

References

1. Choi, S. & Wan, L. The rise of service robots in the hospitality industry: some actionable insights. *Boston Hospitality Rev.* 1–11 (2021).
2. Patil, D. et al. A survey on autonomous military service robot. in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 1–7 (IEEE, 2020). https://doi.org/10.1109/ic-ETITE47903.2020.78
3. Hambuchen, K., Marquez, J. & Fong, T. A review of NASA human-robot interaction in space. *Curr. Rob. Rep.* **2**, 265–272 (2021).
4. Savage, N. Robots rise to meet the challenge of caring for old people. *Nature* **601**, 8–10 (2022).
5. Bushweller, K. Teachers, the robots are coming. But that's not a bad thing. *Education Week* **7**, (2020).
6. Krueger, F., Mitchell, K. C., Deshpande, G. & Katz, J. S. Human–dog relationships as a working framework for exploring human–robot attachment: a multidisciplinary review. *Anim. Cogn.* **24**, 371–385 (2021).
7. Rothstein, N. J., Connolly, D. H., de Visser, E. J. & Phillips, E. Perceptions of infidelity with sex robots. in Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI) 129–139 (2021).
8. Phillips, Elizabeth, Kristin E. Schaefer, Deborah R. Billings, Florian Jentsch, and Peter A. Hancock. Human-animal teams as an analog for future human-robot teams: Influencing design and fostering trust. *J Human-Robot Interact.* **5**(1): 100-125 (2016).
9. de Visser, E. & Parasuraman, R. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *J. Cogn. Eng. Decis. Mak.* **5**, 209–231 (2011).
10. Wickens, C. D., Dixon, S. R. & Ambinder, M. S. Workload and automation reliability in unmanned air vehicles. in Human Factors of Remotely Operated Vehicles (eds Cooke, N. J., Pringle, H. L., Pedersen, H. K. & Connor, O.) vol. 7 209–222 (Emerald Group, 2006).
11. Malle, B. F. & Ullman, D. A multidimensional conception and measure of human-robot trust. in Trust in human-robot Interaction: Research and Applications (eds Nam, C. S. & Lyons, J. B.) 3–25 (Academic Press, 2021).
12. Sharkey, A. Can we program or train robots to be good? *Ethics Inf. Technol.* **22**, 283–295 (2020).
13. van Wynsberghe, A. & Robbins, S. Critiquing the reasons for making artificial moral agents. *Sci Eng. Ethics.* **25**, 719–735 (2019).
14. Russell, S., Aguirre, A., Javorsky, E. & Tegmark, M. Lethal autonomous weapons exist; they must be banned. *IEEE Spectr.* (2021).
15. Denecke, K. & Baudoin, C. R. A review of artificial intelligence and robotics in transformed health ecosystems. *Front. Med.* **9**, 795957 (2022).

16. Chang, C. & Murphy, R. R. Towards robot-assisted mass-casualty triage. in Proceedings of the IEEE International Conference on Networking, Sensing and Control 267–272 (IEEE, 2007).
17. Gogoshin, D. L. Robot responsibility and moral community. *Front. Rob. AI.* **8**, 768092 (2021).
18. Malle, B. F., Bello, P. & Scheutz, M. Requirements for an artificial agent with norm competence. in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society 21–27 (2019). https://doi.org/10.1145/3306618.3314252
19. Bicchieri, C. *The Grammar of Society: the Nature and Dynamics of Social Norms* (Cambridge University Press, 2005).
20. Brennan, G., Eriksson, L., Goodin, R. E. & Southwood, N. *Explaining Norms* (Oxford University Press, 2013).
21. Malle, B. F. What are norms and how is norm compliance regulated? in Motivation and Morality: A Biopsychosocial Approach. (eds Berg, M. & Chang, E. C.) 46–75 (American Psychological Association, 2023).
22. Malle, B. F., Scheutz, M. & Austerweil, J. L. Networks of social and moral norms in human and robot agents. in A World with Robots: International Conference on Robot Ethics: ICRE 2015 (eds Aldinhas Ferreira, M. I., Silva Sequeira, J., Tokhi, M. O., E. Kadar, E. & Virk, G. S.) 3–17 (Springer, 2017).
23. Harsanyi, J. C. Morality and the theory of rational behavior. *Soc. Res.* **44,** 623–656 (1977).
24. Elster, J. Social norms and economic theory. *J. Economic Perspect.* **3**, 99–117 (1989).
25. Krupka, E. & Weber, R. A. The focusing and informational effects of norms on pro-social behavior. *J. Econ. Psychol.* **30**, 307–320 (2009).
26. Postlewaite, A. Social norms and social assets. *Annual Rev. Econ.* **3**, 239–259 (2011).
27. Sunstein, C. R. Social norms and social roles. *Columbia Law Rev.* **96**, 903–968 (1996).
28. Lessig, L. Social meaning and social norms. *Univ. Pa. Law Rev.* **144**, 2181–2189 (1996).
29. Patterson, O. Making sense of culture. *Ann. Rev. Sociol.* **40**, 1–30 (2014).
30. Awad, E. et al. The moral machine experiment. *Nature* **563**, 59–64 (2018).
31. Laakasuo, M. et al. Moral psychology of nursing robots: exploring the role of robots in dilemmas of patient autonomy. *Eur. J. Social Psychol.* **53**, 108–128 (2023).
32. Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J. & Cusimano, C. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI) 117–124 (ACM, 2015).
33. Malle, B. F., Thapa, S. & Scheutz, M. AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. in Robotics and Well-Being (eds Aldinhas Ferreira, M. I., Silva Sequeira, J., Singh Virk, G., Tokhi, M. O. & E. Kadar, E.) 111–133 (Springer International, 2019). https://doi.org/10.1007/978-3-030-12524-0_11
34. Demaree-Cotton, J. & Kahane, G. Moral dilemmas. in Cambridge Handbook of Moral Psychology (eds Malle, B. F. & Robbins, P.) 101–123 (Cambridge University Press, 2025).
35. Dahl, A., Gingo, M., Uttich, K. & Turiel, E. Moral reasoning about human welfare in adolescents and adults: judging conflicts involving sacrificing and saving lives: I. Introduction. *Monogr. Soc. Res. Child Dev.* **83**, 7–30 (2018).
36. Everett, J. A., Pizarro, D. A. & Crockett, M. J. Inference of trustworthiness from intuitive moral judgments. *J. Exp. Psychol. Gen.* **145**, 772 (2016).
37. Malle, B. F. & Phillips, E. A robot's justifications, but not explanations, mitigate people's moral criticism and preserve their trust. *Preprint at* https://doi.org/10.31234/osf.io/dzvn4 (2023).
38. Felzmann, H., Fosch-Villaronga, E., Lutz, C. & Tamo-Larrieux, A. Robots and transparency: the multiple dimensions of transparency in the context of robot technologies. *IEEE Rob. Autom. Magazine.* **26**, 71–78 (2019).
39. Matthews, G., Lin, J., Panganiban, A. R. & Long, M. D. Individual differences in trust in autonomous robots: implications for transparency. *IEEE Trans. Human-Machine Syst.* **50**, 234–244 (2020).
40. Nesset, B., Robb, D., Lopes, J. D. Á. & Hastie, H. Transparency in HRI: trust and decision making in the face of robot errors in Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI) (ACM, 2021).
41. Ososky, S., Sanders, T., Jentsch, F., Hancock, P. & Chen, J. Y. C. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. in Unmanned Systems Technology XVI vol. 9084 112–123 (SPIE, 2014).
42. Fleischmann, K. R. & Wallace, W. A. A covenant with transparency: opening the black box of models. *Commun. ACM - Adapt. Complex. Enterprises.* **48**, 93–97 (2005).
43. Winfield, A. F. et al. IEEE P7001: A proposed standard on transparency. *Front. Rob. AI.* **8**, 665729 (2021).
44. Spagnolli, A., Frank, L. E., Haselager, P. & Kirsh, D. Transparency as an ethical safeguard. in Symbiotic Interaction: 6th International Workshop, Symbiotic 2017, Revised Selected Papers (eds Ham, J. et al) 1–6 (Springer, 2018).
45. Hilton, D. J. & Slugoski, B. R. Knowledge-based causal attribution: the abnormal conditions focus model. *Psychol. Rev.* **93**, 75–88 (1986).
46. Malle, B. F. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction* (MIT Press, 2004).
47. Winikoff, M., Sidorenko, G., Dignum, V. & Dignum, F. Why bad coffee? Explaining BDI agent behaviour with valuings. *Artif. Intell.* **300**, 103554 (2021).
48. Edmonds, M. et al. A Tale of two explanations: enhancing human trust by explaining robot behavior. *Sci. Rob.* **4**, eaay4663 (2019).
49. Miller, T. Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019).
50. Stange, S. & Kopp, S. Effects of a social robot's self-explanations on how humans understand and evaluate its behavior. in Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI) 619–627 (ACM, 2020). https://doi.org/10.1145/3319502.3374802
51. Das, D., Banerjee, S. & Chernova, S. Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery. in Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI) 351–360 (ACM, 2021). https://doi.org/10.1145/3434073.3444657
52. Du, N. et al. Look who's talking now: implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transp. Res. Part. C: Emerg. Technol.* **104**, 428–442 (2019).
53. Tolmeijer, S. et al. Taxonomy of trust-relevant failures and mitigation strategies. in Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI) 3–12 (2020). https://doi.org/10.1145/3319502.3374793
54. Esterwood, C. & Robert, L. P. A literature review of trust repair in HRI. in Proceedings of the 31st IEEE international conference on robot and human interactive communication (RO-MAN) 1641–1646 (IEEE, 2022).
55. Wagner, A. R. & Robinette, P. An explanation is not an excuse: Trust calibration in an age of transparent robots. in Trust in Human-Robot Interaction (eds Nam, C. S. & Lyons, J. B.) 197–208 (Academic Press, 2021).
56. Hermann, J. Moral justification. In *On Moral Certainty, Justification and Practice: A Wittgensteinian Perspective 67–85* (Palgrave Macmillan UK, 2015).
57. Scheutz, M., Malle, B. F. & Briggs, G. Towards morally sensitive action selection for autonomous social robots. in Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) 492–497 (IEEE Press, 2015). https://doi.org/10.1109/ROMAN.2015.7333661
58. Setman, S. A. A willingness to be vulnerable: norm psychology and human–robot relationships. *Ethics Inf. Technol.* **23**, 815–824 (2021).
59. Kasenberg, D., Roque, A., Thielstrom, R., Chita-Tegmark, M. & Scheutz, M. Generating justifications for norm-related agent decisions. in Proceedings of the 12th International Conference on Natural Language Generation (eds van Deemter, K., Lin, C. & Takamura, H.) 484–493 (2019). https://doi.org/10.18653/v1/W19-8660

60. Luebbers, M. B., Tabrez, A., Ruvane, K. & Hayes, B. Autonomous Justification for Enabling Explainable Decision Support in Human-Robot Teaming. in Proceedings of Robotics: Science and Systems (2023). https://doi.org/10.15607/RSS.2023.XIX.002
61. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.
62. Parasuraman, R. & Miller, C. A. Trust and etiquette in high-criticality automated systems. *Commun. ACM*. **47**, 51–55 (2004).
63. Ullman, D. & Malle, B. F. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. in ACM/IEEE International Conference on Human-Robot Interaction (HRI) 618–619 (IEEE, 2019).
64. Law, T., Scheutz, M. & Trust Recent concepts and evaluations in human-robot interaction. in Trust in Human-Robot Interaction (eds Nam, C. S. & Lyons, J. B.) 27–57 (Academic Press, 2021).
65. Malle, B. F. & Ullman, D. A multidimensional conception and measure of human-robot trust. in Trust in Human-Robot Interaction (eds Nam, C. S. & Lyons, J. B.) 3–25 (Academic Press, 2021).
66. Law, T., Malle, B. F. & Scheutz, M. A touching connection: how observing robotic touch can affect human trust in a robot. *Int. J. Social Robot.* **13**, 2003–2019 (2021).
67. Khavas, Z. R., Kotturu, M. R., Ahmadzadeh, S. R. & Robinette, P. Do humans trust robots that violate moral trust? *ACM Trans. Human-Robot Interact.* **13**, 1–30 (2024).
68. Kvalnes, Ø. Springer International Publishing, Cham,. Moral dilemmas. in Moral Reasoning at Work: Rethinking Ethics in Organizations (ed. Kvalnes, Ø.) 11–19 (2019).
69. McConnell, T. Moral dilemmas. In *The Stanford Encyclopedia of Philosophy* (Sprig 2024 Editon) (eds Zalta, E. N. & Nodelman, U.). https://plato.stanford.edu/archives/spr2024/entries/moral-dilemmas
70. Howe, E. S. Integration of mitigation, intention, and outcome damage information, by students and circuit court judges. *J. Appl. Soc. Psychol.* **21**, 875–895 (1991).
71. Malle, B. F., Guglielmo, S. & Monroe, A. E. A theory of blame. *Psychol. Inq.* **25**, 147–186 (2014).
72. Komatsu, T., Malle, B. F. & Scheutz, M. Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across U.S. and Japan. in Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI) 63–72 (IEEE Press, 2021). https://doi.org/10.1145/3434073.3444672
73. Ullman, D. & Malle, B. F. MDMT (Multi-Dimensional Measure of Trust) v2. (2021). https://research.clps.brown.edu/SocCogSci/Measures/MDMT_v2.pdf
74. Mayer, R. C., Davis, J. H. & Schoorman, F. D. An integrative model of organizational trust. *Acad. Manage. Rev.* **20**, 709–734 (1995).
75. Jian, J. Y., Bisantz, A. M. & Drury, C. G. Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* **4**, 53–71 (2000).
76. Schaefer, K. E. Measuring trust in human robot interactions: development of the "Trust perception scale-HRI". in Robust Intelligence and Trust in Autonomous Systems (eds Mittu, R. et al) 191–218 (Springer, 2016).
77. Bigman, Y. E. & Gray, K. People are averse to machines making moral decisions. *Cognition* **181**, 21–34 (2018).
78. Sharkey, A. Can robots be responsible moral agents? And why should we care? *Connection Sci.* **29**, 210–216 (2017).
79. Pak, R. & Rovira, E. A theoretical model to explain mixed effects of trust repair strategies in autonomous systems. *Theoretical Issues Ergon. Sci.* **25**, 453–473 (2024).
80. De Visser, E. J. et al. Towards a theory of longitudinal trust calibration in human–robot teams. *Int. J. Social Robot.* **12**, 459–478 (2020).
81. Briggs, G., Scheutz, M. & Sorry I can't do that: Developing mechanisms to appropriately reject directives in human-robot interactions. in Papers of the 2015 AAAI Fall Symposium 32-36 (2015). https://cdn.aaai.org/ocs/11709/11709-51307-1-PB.pdf
82. Briggs, G., Williams, T., Jackson, R. B. & Scheutz, M. Why and how robots should say 'no'. *Int. J. Soc. Rob.* **14**, 323–339 (2022).
83. Briggs, G. & Scheutz, M. Investigating the effects of robotic displays of protest and distress. In *Social Robotics* (eds Ge, S. S. et al.) 238–247 (Springer, 2012).
84. Zhu, Q., Williams, T., Jackson, B. & Wen, R. Blame-laden moral rebukes and the morally competent robot: A Confucian ethical perspective. *Sci. Eng. Ethics.* **26**, 2511–2526 (2020).
85. Chi, V. B. & Malle, B. F. People dynamically update trust when interactively teaching robots. in Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI) 554–564 (2023). https://doi.org/10.1145/3568162.3576962
86. Petty, R. E. & Cacioppo, J. T. The elaboration likelihood model of persuasion. in *Advances in Experimental Social Psychology* (ed Berkowitz, L.) vol 19 123–205 (Academic Press, 1986).
87. Doyle-Burke, D. & Haring, K. S. Robots are moral actors: unpacking current moral HRI research through a moral foundations lens. In *Social Robotics* (eds Wagner, A. R. et al.) 170–181 (Springer International, 2020).
88. Chita-Tegmark, M., Law, T., Rabb, N. & Scheutz, M. Can you trust your trust measure? in Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI) 92–100 (2021). https://doi.org/10.1145/3434073.3444677

Acknowledgements

This work is supported by the U.S. Air Force Office of Scientific Research award number FA9550-21-1-0359. The views expressed in this paper are those of the authors and do not reflect those of the U.S. Air Force, U.S. Department of Defense, or U.S. Government.

Author contributions

Authors EP and BFM jointly designed and conducted the reported research, analyzed the data, and cowrote the main manuscript including tables, as well as the supplementary materials. Author BFM prepared all figures in the main manuscript and in the supplementary materials. All authors reviewed the manuscript before submission.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-17983-2.

Correspondence and requests for materials should be addressed to E.K.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.