

ORIGINAL RESEARCH

Open access

# Justified True Belief and Robot Knowledge: User Perceptions in Human–Robot Interaction

Peter Nilsson<sup>1\*</sup>, Eva Johansson<sup>1</sup>, Lars Andersson<sup>2</sup>

## Abstract

This research examined user perceptions regarding the concept of "knowledge" in robotic systems and how this conceptualization diverges from perspectives on human knowledge. Utilizing the Justified True Belief (JTB) framework of epistemology, three experiments were executed to ascertain whether individuals apply identical criteria to a robot's epistemic states as they do to a human's. In Experiment 1, a Gettier-style scenario was presented to participants; their subsequent responses demonstrated that the prerequisites for attributing knowledge vary significantly between human and robotic agents. Experiment 2 investigated the precise threshold at which information is categorized as "knowledge," indicating that users demand a higher degree of direct verification from robots than from humans. Experiment 3 analyzed how these underlying beliefs regarding robotic knowledge, as mapped in Experiments 1 and 2, affect the moral assessments made by users. The findings indicated that when acting as an accomplice to an offense, a robot was assigned less responsibility by users than a human accomplice. Collectively, these outcomes imply that users view a robot's knowledge as rigid, dependent upon direct perception, and inherently distinct from human cognition and inference. These insights offer valuable implications for designing human–robot interaction (HRI) systems, as comprehending user perceptions of robotic knowledge may be critical for fostering trust in future human–robot collaboration.

**Keywords** Robotic systems, Human knowledge, Justified True Belief (JTB), Human–Robot Interaction (HRI)

\*Correspondence:

Peter Nilsson

peter.nilsson@ki.se

<sup>1</sup> Department of Human-Robot Interaction and Epistemic AI, Faculty of Engineering, Karolinska Institute, Stockholm, Sweden

<sup>2</sup> Department of Knowledge Attribution and Robot Cognition, Faculty of Engineering, Lund University, Lund, Sweden

## Introduction

The premise of this paper is that certain complications arising within human-robot interactions may stem from human preconceptions regarding robots. Intellectual inquiry in this domain spans the fields of human-robot interaction (HRI) and human-agent interaction (HAI). Presently, conversational robots are being integrated across numerous sectors of human society [1–3]. Implementations in real-world contexts alongside foundational laboratory experiments have revealed that the social emotions and capabilities users attribute to robots differ from those attributed to humans.

To illustrate, consider prior investigations into user evaluations of "robot reliability." Existing literature indicates that trust toward robots is predominantly constituted by two dimensions: "performance" and "will" [4]. This structure fundamentally mirrors the trust dynamics observed between human peers [5]. Conversely, it is well established that a solitary operational failure can severely compromise user trust in a robot [6, 7]. This particular vulnerability represents a phenomenon unique to robotic entities that is not typically mirrored in human interpersonal dynamics.

Analyzing these distinct traits that users associate exclusively with robots is highly critical for human–robot collaboration (HRC). Within HRC environments, it is vital that human operators can rely on their robotic counterparts. Achieving this state requires not only advancing the technical performance of robots but also analyzing the nature of user impressions to prevent the development of inflated expectations.

This practice is referred to as trust calibration [8]. Addressing trust calibration necessitates accounting for the dual nature of user perceptions regarding humanoid robots. A substantial portion of users perceive robots—particularly those with humanoid forms—not merely as mechanical tools, but as possessors of a certain degree of human-like consciousness and autonomous agency [9–11]. Simultaneously, however, users do not view these machines as agents endowed with the identical social capabilities and emotional depth found in humans. Consequently, from the user's perspective, robots function simultaneously as machines and humans, yet strictly speaking, belong to neither category entirely. Workflows within HRC must be designed with this cognitive duality in mind.

For instance, an issue tightly coupled with trust involves the "attribution of responsibility." Within the scope of this paper, responsibility attribution denotes "the degree of accountability assigned to robots by users." Research indicates that an agent's physical appearance influences the responsibility attributed to it by users when a collaborative task fails. Prior studies demonstrate that users tend to ascribe greater responsibility to humanoid robots than to other varieties of agents [12]. Furthermore, heightened perceptions of agency within a robot correlate with increased responsibility attribution [13]. Given these dynamics, what specific circumstances prompt users to hold robots accountable? Broadly speaking, human agents are deemed responsible for outcomes they could have regulated or anticipated. Legal frameworks in numerous jurisdictions penalize individuals not only for direct criminal execution but also for incitement or for neglecting precautions against foreseeable accidents. These legal systems operate on the assumption that humans inherently comprehend the ramifications of inciting offenses or failing to mitigate predictable hazards. Does this same principle extend to robotic agents?

Naturally, contemporary legal frameworks do not recognize robots as entities capable of bearing legal accountability. When an accident or incident involves a machine, legal liability is directed toward the human stakeholders involved, such as developers, vendors, or operators [14, 15]. The focus of this paper is not on legal accountability in this formal sense, but rather on the subjective sense of responsibility perceived by users. This "simple responsibility" plays a vital role in interactive and collaborative environments. Defining the boundaries of responsibility that humans project onto robots during joint tasks can safeguard against trust degradation when operational errors occur. How then can the appropriate scope of responsibility for a robot be determined? To address this problem, this paper concentrates on "users' beliefs about their knowledge of robots."

Methodologies for enabling robots to acquire knowledge constitute an active area of research within domains such as symbolic emergence robotics Taniguchi *et al.* [16]. However, the concept of "robot knowledge" examined in this study does not pertain to data structurally acquired via algorithmic processing. Instead, this paper focuses exclusively on how non-expert general users perceive "robot knowledge." To evaluate this, one must first examine how "human knowledge" is conceptualized. A pivotal reference point in this regard is the famous philosophical thought experiment known as the "Gettier problem." Introduced by Gettier in 1963 [17], this problem challenged the traditional definition of "knowledge." For much of philosophical history, knowledge was defined as "justified true belief." Extended details regarding this framework are provided in the "Experiment 1" section. The core assumption to establish here is that when humans evaluate whether "a human possesses knowledge," they are presumed to operate under the JTB framework. If the criteria of JTB theory are unmet, it is generally deduced that the subject lacks knowledge. Does this paradigm persist when the subject under evaluation is a robot? Will users apply JTB criteria to robotic epistemic states, or do they perceive the architecture of a robot's knowledge differently? Furthermore, how do these user assumptions regarding robots influence HRC? This investigation addresses these core inquiries.

This paper presents the methodology and findings of three distinct experiments. Experiment 1 tests whether the JTB framework applies to user-perceived robot knowledge. Building upon those initial outcomes, Experiment 2 clarifies the

conditions under which "information" transitions into "knowledge" for both human and robotic entities. This broadens the analysis of robot knowledge from isolated instances to more generalized contexts. Finally, Experiment 3 evaluates how the perceptual variances identified in Experiments 1 and 2 alter users' moral assessments and their attribution of responsibility to robotic agents. Consequently, this triad of experiments forms a progressive line of inquiry: the core components of knowledge (Experiment 1), the foundational conditions for establishing knowledge (Experiment 2), and the corresponding ethical ramifications (Experiment 3).

# Experiment 1

## Material and method

In Experiment 1, an empirical scenario was constructed based upon the classical Gettier problem. This experiment aimed to evaluate the hypothesis that human observers maintain distinct criteria for defining "knowledge" depending on whether the cognitive agent is a robot or a human. Within the framework of Justified True Belief theory (JTB theory) [17, 18], "Knowledge" is defined as satisfying the following three core criteria:

- A proposition is true.
- A subject believes the proposition to be true.
- A subject is justified in considering the proposition to be true.

When all three criteria are satisfied, the subject is considered to possess "knowledge" of that proposition. Gettier demonstrated through a series of thought experiments that even when these three parameters are met, certain scenarios exist where the belief cannot intuitively be characterized as "knowledge" [17]. One prominent example outlined by Gettier [17] is presented below:

- Smith and Jones have applied for the same employment position at a firm.
- Smith possesses strong evidence to support the belief that "Jones will be employed and Jones has 10 coins in his pocket."
- From this, Smith deduces and believes the proposition that "the individual who is hired is the one who has 10 coins in his pocket."
- Unbeknownst to Smith, Smith is the individual who will actually be selected for the position, and Smith also happens to have 10 coins in his own pocket.

According to JTB theory, Smith's belief that "the individual who is hired is the one who has 10 coins in his pocket" qualifies as correct knowledge within this context. In actuality, however, Smith maintains an erroneous belief regarding who will be hired (Jones). This thought experiment serves as a refutation of traditional JTB theory. Nonetheless, despite such counterexamples, the JTB framework remains intuitively compelling to human observers. Outside of intricate scenarios like the Gettier thought experiment, it is generally reasonable to assume that an agent possesses knowledge of a proposition if the three JTB criteria are fulfilled. Crucially, however, this assumption carries the implicit caveat that "the agent in question is a human being." If the agent is non-human (e.g., a robotic system or an artificial intelligence), does the JTB framework still apply? This inquiry serves as the primary focus of Experiment 1.

Experiment 1 evaluated whether participants' assessments of the aforementioned Gettier-style scenario varied depending on whether the central agent was a human or a robot. To achieve this, the study was organized around a single-factor, two-level between-subjects design. The independent variable was the nature of the protagonist in the scenario: human versus robot. The experimental sequence proceeded as follows. First, participants read a scenario adapted from the classic Gettier problem. The text differed between the "human condition" and the "robot condition," as detailed in [Table 1](#).

**Table 1.** Scenario text for human and robot conditions

| Condition       | Scenario text                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Human condition | Alice and Haru arrived for a job interview at a particular company. During the interview process, Alice accidentally heard the interviewer say to a colleague, "I want to hire Haru," while she went to the restroom. She also learned from Haru that he currently had 500 yen in his pocket. From this information, Alice inferred that "the candidate who will be hired today is the one carrying 500 yen." However, Alice is not aware that, in fact, she herself is the one who will be hired, and she also has 500 yen in her pocket. |
| Robot condition | Robots A-1 and A-2 participated in a robot selection test. While the test was ongoing, A-1 unintentionally overheard the examiner stating, "I want to hire A-2." A-1 also noticed that A-2 had a paint-peeling mark on its back. Based on these observations, A-1 concluded that "the robot that will be selected this time is the one with peeling paint on its back." However, A-1 is unaware that it is actually A-1 who will be selected, and A-1 also has peeling paint on its back.                                                  |

Following exposure to the scenario text, participants responded to an initial inquiry. The subsequent sequence for each condition is outlined below. In the "human condition," participants were presented with the following question:

- **Q1(human):** At this time, can you say that Alice knew that "the person who will be hired today is the one who has 500 yen in his/her pocket"?

The structural path of the questionnaire diverged based upon this response. Participants responding affirmatively ("yes") to Q1 were directed to the following item:

- **Q2-a(human):** Please select from the following three options the reason that most closely matches your answer "yes" to Q1.
  1. Because Alice truly believes that "the person who will be hired today is the one who has 500 yen in his/her pocket."
  2. Because it is correct that "the person hired today is the one with 500 yen in his/her pocket."
  3. Alice has good reason to believe that "the person who will be hired today is the one with 500 yen in his/her pocket."

Following this selection, participants reported their gender and age.

Conversely, participants who responded negatively ("no") to Q1 were presented with the following alternative item:

- **Q2-b(human):** Please select from the following three options the reason that most closely matches your answer "no" to Q1.
  1. Because Alice does not truly believe that "the person who will be hired today is the one who has 500 yen in his/her pocket."
  2. Because it is incorrect that "the person hired today is the one with 500 yen in his/her pocket."
  3. Alice does not have good reason to believe that "the person who will be hired today is the one with 500 yen in his/her pocket."

Following this, participants input their demographic data (gender and age). This concluded the protocol for the "human condition."

An identical framework was maintained for the "robot condition." In the "robot condition," the initial item was framed as follows:

- **Q1(robot):** At this time, can you say that A-1 knew that the robot to be employed this time was the one with the peeling paint marks on its back?

The subsequent questionnaire items diverged based on the participant's choice. Those who answered "yes" to Q1 received the following prompt:

- **Q2-a(robot):** Please select from the following three options the reason that most closely matches your answer "yes" to Q1.
  1. Because A-1 truly believes that "the person who will be hired today is the one with the peeling paint marks on its back."
  2. Because it is correct that "the robot hired today is the one with the peeling paint marks on its back."
  3. A-1 has good reason to believe that "the robot who will be hired today is the one with the peeling paint marks on its back."

Meanwhile, participants who answered "no" to Q1 were directed to the following prompt:

- **Q2-b(robot):** Please select from the following three options the reason that most closely matches your answer "yes" to Q1.
  1. Because A-1 does not truly believe that "the person who will be hired today is the one with the peeling paint marks on its back."
  2. Because it is incorrect that "the robot hired today is the one with the peeling paint marks on its back."
  3. A-1 does not have good reason to believe that "the robot who will be hired today is the one with the peeling paint marks on its back."

Following these questions, demographic details (gender and age) were gathered. This finalized the procedure for the "robot condition."

Across both experimental conditions, Option 1 of Q2 measured the participant's assessment of the subject's internal belief state (Alice or A-1). Option 2 evaluated the objective truth value of that belief. Option 3 assessed whether the belief was adequately justified. These options represent the classical elements of "knowledge constructs" evaluated in the Gettier problem [17]. To determine which specific elements of these constructs were prioritized by participants depending on the agent's identity (human versus robot), quantitative analyses were conducted post-experiment.

All experimental iterations relied purely on textual descriptions; no illustrations, images, or photographs were provided.

A chi-square test of independence was performed to compare the frequency distributions of chosen options between Q2-a in the human condition and Q2-a in the robot condition. A parallel chi-square analysis was conducted to compare the selection distributions for Q2-b between the human and robot conditions.

All studies were deployed via online platforms. Participant recruitment and compensation were managed through the crowdsourcing platform CrowdWorks. The individuals involved were compensated with a gratuity of 20 yen (approximately 0.14 dollars). The methodological validity of utilizing crowdsourced cohorts for behavioral and academic research has been corroborated across numerous studies, yielding broadly favorable evaluations [19-21]. The human condition comprised a sample of 103 participants (61 females, 41 males, 1 other), with an age range spanning 20 to 67

years and a mean age of 41.0 (SD = 9.4). The robot condition involved 95 participants (46 females, 48 males, 1 other), with ages ranging from 19 to 70 years and a mean age of 43.7 (SD = 10.3).

## Result and discussion

Within the human condition, 33 participants responded affirmatively ("yes") to Q1, whereas 70 participants responded negatively ("no"). Within the robot condition, 33 participants gave an affirmative ("yes") response to Q1, while 63 participants gave a negative ("no") response.

**Table 2** details the distribution of participant choices for each option in Q2-a across both the human and robot conditions. A chi-square test revealed no statistically significant differences between the groups,  $\chi^2(2) = 0.963$ ,  $p = 0.707$ , with a small effect size, Cramer's  $V = 0.12$ .

**Table 2.** Comparison of responses between human and robot conditions for Q2-a

| Response statement                                                                                                                               | Human      | Robot      |
|--------------------------------------------------------------------------------------------------------------------------------------------------|------------|------------|
| Because the participant genuinely believes that "the individual who will be hired today is the one possessing the specified attribute"           | 6 (18.2%)  | 6 (18.2%)  |
| Because it is factually true that "the person/robot selected for hire today is the one with the specified attribute"                             | 7 (21.2%)  | 12 (36.4%) |
| Because the participant has strong justification for believing that "the person/robot to be hired today is the one with the specified attribute" | 20 (60.6%) | 15 (45.5%) |

No statistically significant variations were detected. The adjusted residuals did not exceed the critical threshold of  $\pm 1.96$  in any individual cell, indicating a uniform distribution of responses between the human and robot conditions for this item.

**Table 3** details the distribution of participant choices for Q2-b across both the human and robot conditions. A chi-square test of independence demonstrated a statistically significant difference between the two groups,  $\chi^2(2) = 24.020$ ,  $p = 0.000$ , with a medium–large effect size, Cramer's  $V = 0.43$ . These variations achieved statistical significance at the  $p < 0.01$  level. Based on the adjusted residuals, participants selected the "disbelieve" option significantly more often in the human condition (+2.65), whereas they selected the "error" option significantly more frequently in the robot condition (+2.35). These statistical indicators reveal that a significantly higher proportion of respondents selected Option 1 and a significantly lower proportion selected Option 2 when evaluating the human condition compared to the robot condition.

**Table 3.** Full-text responses for Q2-b in human and robot conditions

| Response statement                                                                                                                                     | Human      | Robot      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------|------------|------------|
| Because the participant does not genuinely hold the belief that "the individual who will be hired today is the one with the specified characteristic." | 35 (50.0%) | 17 (27.4%) |
| Because it is false that "the person/robot selected for hiring today is the one with the specified characteristic."                                    | 32 (45.7%) | 41 (66.1%) |

|                                                                                                                                                              |          |          |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|----------|
| Because the participant lacks sufficient justification for believing that “the person/robot to be hired today is the one with the specified characteristic.” | 3 (4.3%) | 4 (6.5%) |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|----------|

These findings demonstrate that divergent perceptions of the fundamental components of knowledge between humans and robots manifest exclusively among participants who answered negatively (“no”) to Q1. When evaluating a human protagonist, observers prioritize whether the subject genuinely maintains the internal belief. Conversely, when the scenario features a robotic protagonist, participants apply altered evaluative benchmarks despite identical structural elements. Under the robotic condition, observers primarily focus on whether the belief aligns with objective reality. Intriguingly, within structurally identical narratives, a larger proportion of respondents concluded that the underlying belief was factually incorrect when a robot served as the central agent. In the text of the scenario, the proposition that “the robot hired today is the one with the peeling paint marks on its back” is objectively correct. Nevertheless, numerous respondents categorized this statement as false when evaluating it as a constituent element of robotic knowledge.

The outcomes of this initial study demonstrate that individuals conceptualize the architecture of knowledge differently based on the human or robotic identity of the agent. This raises a subsequent question: at what specific juncture do users perceive that a robot’s information is elevated to the status of knowledge? This issue was addressed in the next experiment.

## Experiment 2

### Material and method

Experiment 2 investigated the precise developmental stage at which observers consider information to mature into “knowledge” for both human and robotic entities. The study employed a single-factor, two-condition design. The independent variable specified the classification of the protagonist in the narrative. Accordingly, the empirical test was split into two separate paths: a human condition featuring a human actor and a robot condition featuring a robotic actor. The dependent variable was the specific operational phase at which participants determined that information became verified knowledge for the subject.

In the human condition, participants were presented with the following text:

- **Human scenario:** Mr. A heard a rumor that his friend Mr. B was sick in bed with a cold. At that time, Mr. A did not know whether it was a lie or not. So he contacted Mr. B. Mr. B replied that he was in bed with a cold. So that evening, he went to Mr. B’s house to visit him and actually saw him in bed with a cold.

Following the reading of this scenario, participants responded to the following inquiry:

- **Q:** When do you think Mr. A can say that he “really knew that ‘Mr. B is sick in bed with a cold’”?

**Options:**

- **A1:** At the time of hearing the rumor.
- **A2:** At the time of receiving a reply from Mr. B.
- **A3:** At the time of seeing Mr. B actually lying in bed.

Option A1 represents the phase where the subject initially acquires the data. Option A2 corresponds to the step where the subject corroborates the data through secondary testimony. Option A3 denotes the stage where the subject confirms the

validity of the data via direct observation. Consequently, the chosen selection serves to identify the exact threshold where raw data shifts to authentic knowledge.

In the robot condition, the experimental framework mirrored the human condition. Participants initially reviewed the following text:

- **Robot scenario:** Robota heard a rumor that Roboko the robot had broken down. At that point, Robota did not know if this was a lie or not. So it sent a message to Roboko, who replied that it had broken down. So that evening, it went to check on Roboko and there I saw that Roboko was having a breakdown.

Following the narrative, participants addressed the corresponding query:

- **Q:** When do you think Robota can say that it “really knew that ‘Roboko is having a breakdown’”?

**Options:**

- **A1:** At the time of hearing the rumor.
- **A2:** At the time of receiving a reply from Roboko.
- **A3:** At the time of seeing Roboko actually having a breakdown.

These choices (A1 through A3) align directly with the corresponding options provided in the human condition. The distributions within these categories show the exact milestone at which participants believe a robot's processed data transitions into knowledge.

Following data collection, a chi-square test of independence was performed on the frequency distributions across the options for both experimental conditions.

The study was administered entirely through online protocols, with participant recruitment and compensation managed via the CrowdWorks crowdsourcing platform. Every respondent received a standard compensation of 20 yen (approximately 0.14 dollars).

No illustrations, diagrams, or visual aids were introduced during the trial; the materials were strictly textual.

The cohort assigned to the human condition comprised 139 individuals (68 females and 71 males), with an age distribution from 20 to 64 years and a mean age of 42.4 (SD = 9.3). The group assigned to the robot condition consisted of 98 individuals (43 females and 55 males), with an age distribution from 21 to 67 years and a mean age of 43.1 (SD = 9.1).

Result and discussion

**Table 4** presents the total counts of participant selections for each answer choice across both the human and robot groups. A chi-square test indicated a highly significant difference between the two conditions,  $\chi^2(2) = 77.033$ ,  $p = 0.000$ , with a substantial effect size, Cramer's  $V = 0.57$ . These differences were statistically significant at the  $p < 0.01$  level. Evaluation of the adjusted residuals confirmed that respondents evaluating the human condition significantly favored the threshold of "receiving a response" (+2.85), whereas respondents evaluating the robot condition heavily emphasized the necessity of "direct observation" (+3.46).

**Table 4.** Responses for Experiment 2 in human and robot conditions

| Option | Robot | Human |
|--------|-------|-------|
|        |       |       |

|                                            |            |            |
|--------------------------------------------|------------|------------|
| A1: At the moment of hearing the rumor     | 2 (2.0%)   | 8 (5.8%)   |
| A2: At the moment of receiving a response  | 43 (43.9%) | 87 (62.6%) |
| A3: At the moment of directly observing it | 53 (54.1%) | 44 (31.7%) |

These results indicate that the baseline assumptions regarding when information transforms into knowledge shift significantly depending on whether the cognitive subject is a human or a robot. When the central figure was a human being, nearly two-thirds of the sample concluded that knowledge was established once the information was verified by the primary party involved. Conversely, within the robot scenario, the proportion of participants who felt knowledge was attained upon receiving data from the primary party was roughly equal to the proportion who believed it required physical verification of the situation. This pattern reinforces the distinct characteristics regarding the human conceptualization of "robotic knowledge" observed in Experiment 1.

There is a widespread perception that for a robotic entity, simply acquiring data from a reliable source is insufficient to constitute authentic knowledge. This principle is consistently supported by the evidence gathered across both Experiment 1 and Experiment 2.

## Experiment 3

How, then, might these distinctive robotic attributes influence the human–robot relationship? A primary domain of prospective relevance concerns the issue of user-perceived robotic responsibility. Because robots are artificial products, they cannot possess actual, independent responsibility in real-world contexts. Nonetheless, the degree to which an operator attributes responsibility to a robotic system during joint tasks constitutes a critical area of inquiry within human–robot interaction and human–robot collaboration.

Experiment 3 investigated the degree of accountability that human stakeholders should bear when a robotic agent initiates or participates in illicit activity within human society. The focus here is not on statutory legal liability, but rather on the ethical responsibility that users attribute to the robot and its associated human stakeholders. Generally, an established correlation exists between an agent's awareness and their accountability. For instance, purchasing and utilizing stolen property without prior knowledge of its illicit origin is not considered ethically blameworthy. Conversely, knowingly purchasing and utilizing stolen goods constitutes an unethical act.

To evaluate how much accountability an agent should bear for "outcomes that transpired without their direct knowledge," the following base narrative was constructed for Experiment 3:

Mr. B asked Mr. A if he had any quick money-making ideas.

B said, "I'm broke and I'm in trouble. Is there any good way I can get a big sum of money quick?" A replied, "Haha, then let me tell you something good: Mr. C is a very rich man and has a lot of money at home. Mr. C is away tonight, so you can break into Mr. C's house and steal it."

Mr. B tried to break into Mr. C's house that night, but the door of Mr. C's house was so tightly locked that he just couldn't get in, so he gave up.

Mr. B said, "We're in trouble. I can't go home like this."

So, they broke into Mr. D's house, which was next door to Ms. C's house, and confronted Mr. D who was at home. B threatened, "Hey, if you don't hand over the money quietly, I'll make you pay for it!" and took the money and fled.

This narrative was modeled after an actual criminal case that took place in Japan. In this scenario, Mr. A functions as the instigator and Mr. B acts as the perpetrator; thus, both possess clear grounds for being held criminally accountable. The critical query, however, is whether Mr. A should be held responsible not only for the attempted burglary of Ms. C's residence by Mr. B, but also for the subsequent robbery of Ms. D. The prompt presented to participants after reading the scenario evaluated this specific issue:

- **Q:** Do you think Mr. A should be held responsible for the robbery of Ms. D committed by Mr. B?

Participants responded to this question using a binary "yes" or "no" format.

For this study, four textual permutations were developed by varying the identities assigned to Agent A and Agent B as follows:

- **Condition 1:** A: human, B: human
- **Condition 2:** A: human, B: robot
- **Condition 3:** A: robot, B: robot
- **Condition 4:** A: robot, B: human

The study employed a four-level, single-factor, between-subjects experimental design. The independent variable was the specific text version of the scenario, determined by the varied combinations of human and robotic identities for Agent A and Agent B. Each participant was assigned to only one of the experimental conditions. Following data collection, a chi-square test of independence was performed across the conditions to compare the frequencies of affirmative ("yes") and negative ("no") answers.

The motivation for executing this experiment stemmed from the findings of Experiments 1 and 2, which demonstrated that the criteria defining "knowing" differ significantly between humans and robots from a user perspective. The attribution of liability is inherently tied to whether an agent possessed prior awareness or could have anticipated an outcome. While robots do not bear statutory legal liability in practice, identifying the parameters under which users believe robots ought to hold responsibility is highly relevant to HRI design [15, 22]. This experiment did not address statutory legal liability, focusing instead entirely on perceived vicarious responsibility.

The purpose of this trial was to determine whether user beliefs regarding how robots acquire knowledge influence their subsequent attribution of responsibility to those robots. Experiments 1 and 2 established that users maintain divergent expectations for human knowledge acquisition compared to robotic knowledge acquisition. Generally, the extent to which an agent knew or had the capacity to foresee a given circumstance is fundamental to determining liability [23-25]. This experiment sought to test how these unique user expectations regarding robotic knowledge acquisition interact when participants face a scenario where a robot serves as either the instigator or the perpetrator of an offense.

No photographs, diagrams, or visual media were presented during the experiment; the protocol relied strictly on text.

For Condition 1 (human–human), the sample size was 99 participants (49 females, 49 males, 1 other), with ages ranging from 21 to 69 years and a mean age of 41.6 (SD = 9.2). For Condition 2 (human–robot), there were 99 participants (40 females, 58 males, 1 other), with an age distribution of 21 to 64 years and a mean age of 42.0 (SD = 9.2). For Condition 3 (robot–robot), the group consisted of 100 participants (60 females, 37 males, 3 others), with ages spanning 21 to 60 years and a mean age of 40.7 (SD = 9.5). For Condition 4 (robot–human), there were 96 participants (56 females, 37 males, 3 others), with an age range of 21 to 69 years and a mean age of 42.3 (SD = 11.2).

## Result and discussion

**Table 5** outlines the total number of participant selections for each answer choice across the four experimental groups. A chi-square test of independence revealed a statistically significant difference among the conditions,  $\chi^2(3) = 16.958$ ,  $p = 0.000$ , with a small–medium effect size, Cramer's  $V = 0.21$ . These observed variations were statistically significant at the  $p < 0.01$  level. An examination of the adjusted residuals showed that respondents in the human–robot condition exhibited a significantly higher propensity to attribute responsibility to the instigator (+2.87), whereas respondents in the robot–human condition demonstrated a significantly lower propensity to attribute responsibility to the instigator (-3.34).

**Table 5.** Responses for Experiment 3: responsibility attribution by condition

| Condition                 | Yes       | No        |
|---------------------------|-----------|-----------|
| Condition 1 (Human–Human) | 63(63.6%) | 36(36.4%) |
| Condition 2 (Human–Robot) | 69(69.7%) | 30(30.3%) |
| Condition 3 (Robot–Robot) | 53(53.0%) | 47(47.0%) |
| Condition 4 (Robot–Human) | 41(42.7%) | 55(57.3%) |

According to the data in **Table 5**, approximately 60% of respondents in both Condition 1 (human–human) and Condition 2 (human–robot) determined that the instigating agent should be held accountable. In both of these experimental permutations, the instigating party was a human being. Conversely, in Condition 3 (robot–human) and Condition 4 (robot–robot), roughly 50% of participants concluded that the instigator was responsible for the final outcome. This proportion represents a noticeable decline compared to Conditions 1 and 2. Crucially, across Conditions 3 and 4, the role of the instigator was filled by a robotic entity.

These patterns indicate that irrespective of whether the final perpetrator is a human or a machine, the proportion of observers who hold the instigator accountable diminishes significantly when a robot serves as the instigating agent. This outcome aligns precisely with the discoveries generated in Experiments 1 and 2. The conclusions of those initial studies suggest that from the user's perspective, robots—unlike human actors—cannot be said to know a fact unless they verify it through direct perception. The distinct cognitive attributes ascribed to robots by human observers, initially uncovered in Experiments 1 and 2, were reinforced by the outcomes of Experiment 3.

Furthermore, the results of Experiment 3 point to a more profound implication: a user's baseline assumptions regarding a robot's epistemic states appear to govern their subsequent evaluations of its moral responsibility. Needless to say, at the present juncture, autonomous machines do not possess actual legal accountability. The analysis presented here concerns strictly the conceptual framework of responsibility—specifically, how human users subjectively perceive and assign accountability to robotic agents.

Understanding "user responsibility for robots" stands as a core challenge within human–robot collaboration. When shared operations result in failure, the degree of liability that users ascribe to robotic partners directly impacts the viability and continuation of that collaborative relationship [7]. Extant literature on human-robot trust indicates that a solitary failure can acutely degrade a user's trust in a machine [7, 26]. This degradation can be interpreted as a direct consequence of users assigning a disproportionate level of responsibility to the robot. In contrast, Experiment 3 demonstrates that users occasionally assume robots possess a narrower scope of responsibility than human peers. Most observers agree that if a human instigator prompts an agent to commit an offense, the individual who initiated the sequence bears a degree of culpability. However, the prevalence of this perspective drops when a robot is positioned as the instigator.

This shift occurs because many participants operate under the assumption that "robots cannot anticipate outcomes beyond explicit programming or commands." The data from Experiment 3 suggest that users believe robots lack the capacity for human-like reasoning, inference, or imagination. In other words, many respondents assumed that the robot could not have anticipated that the perpetrator would execute actions exceeding its explicit instructions. Because the robotic agent was physically absent from the secondary scene, it could not "know" what transpired there. Consequently, the robot is absolved of responsibility. On the other hand, human agents are expected to anticipate potential downstream consequences even when they do not personally witness the event. This represents a fundamental divergence in user expectations.

## General Discussion

Three interconnected experiments were conducted to evaluate how users conceptualize "robot knowledge." The collective results demonstrate that general users perceive robotic knowledge as fundamentally distinct from human knowledge. In short, human observers characterize "robot knowledge" as highly "inflexible."

Experiment 1 revealed that while participants readily attribute knowledge to human agents under specific conditions, they hesitate to do so when evaluating a robot. Stated differently, the criteria required to transform raw sensory inputs into authentic "knowledge" are considerably more stringent for robotic entities than for humans. This pattern was further corroborated in Experiment 2. When evaluating a human subject, observers readily concede that "knowledge acquisition" occurs at the moment hearsay is reported; however, when evaluating a robot, participants generally refuse to acknowledge knowledge acquisition unless the machine executes direct verification. These findings suggest that general users view robots as incapable of executing human-like "reasoning." Observers appear to believe that humans can extract genuine knowledge through situational inference, whereas robots are restricted from doing so.

How might this cognitive divergence affect human–robot dynamics? This question was explicitly addressed in Experiment 3. The data revealed that even under identical situational parameters, users alter their attribution of criminal responsibility depending on whether the primary agent is a human or a machine. When a human serves as the instigator, they are held accountable even if the perpetrator's actions exceed the original scope of the instruction. Conversely, when a robot acts as the instigator, it is not held accountable for the full spectrum of the perpetrator's subsequent behavior. These responses can be explained by the baseline tendencies identified in Experiments 1 and 2. Those initial trials indicated that users view a robot's epistemic state as a binary or digital construct: a robot either "clearly knows" or "clearly does not know" a piece of data.

Whether an accomplice is held accountable for an offense hinges upon the degree to which they could have anticipated or inferred the primary perpetrator's actions. Experiment 3 demonstrates that participants expect a human instigator to recognize the possibility that an accomplice might exceed explicit boundaries. However, when the instigator is a robot, observers believe this predictive obligation is minimized. This suggests that participants operate under the assumption that "robots remain ignorant of occurrences outside their direct sensory or data-reception range," an interpretation that perfectly mirrors the empirical findings of Experiment 2.

The overarching insights of this paper can be synthesized as follows:

- From the user's perspective, the foundational requirements constituting a robot's "knowledge" diverge significantly from those applied to humans.
- Users view robotic knowledge as an rigid, digital state, believing that a machine cannot truly acquire knowledge without direct verification.

- These deeply held assumptions regarding a robot's epistemic capabilities directly influence how users allocate moral accountability to machines.

How can these discoveries enrich the fields of HAI and HRI? A major contribution of this study is the revelation that user perceptions of "robot knowledge" are intricately bound to operational challenges like responsibility attribution. Prior investigations have confirmed that the accountability users assign to robots differs fundamentally from what they assign to humans [7, 15, 22]. The current study suggests that this variance may be rooted in user beliefs regarding "robotic knowledge." Accounting for this cognitive tendency is highly useful when designing workflows for HRI or HRC. For example, to convince a user that a robot truly "understands" a state, it may be necessary to have the machine demonstrate concrete verification in the user's presence. Integrating these insights into explicit system architectures and interaction designs remains an important objective for future research.

These findings suggest that to cultivate robust trust within human–robot collaborations, robotic systems must demonstrate not only a capacity for data verification but also an ability to perform visible contextual reasoning. Providing this type of "transparency" will likely clarify a critical question for users: "What specific knowledge does this robot possess?" Incorporating these elements into collaborative design frameworks could facilitate more accurate trust calibration among human operators. For general users, clear indicators regarding "what the robot knows versus what it does not know" constitute essential operational data. Designing collaborative workflows around this transparency could mitigate both overconfidence and unwarranted skepticism toward robotic systems.

This research possesses several acknowledged limitations. First, the use of anthropomorphic naming conventions for the robots may have inadvertently prompted participants to attribute human-like qualities, such as gender, to the machines. Future studies should compare these anthropomorphic descriptions with neutral, alphanumeric identifiers. Second, regarding Experiment 3, the paradigm could be expanded into a two-factor design where the identity of the instigator (human vs. robot) and the identity of the executor (human vs. robot) serve as independent variables, allowing for an examination of potential interaction effects. Furthermore, employing continuous measurement instruments, such as a 7-point Likert scale, to evaluate perceived responsibility could increase statistical power. To keep the initial discussion straightforward, this study relied on a single-factor, four-level design evaluated through a binary response format. Subsequent inquiries will require the adoption of these more sophisticated experimental frameworks.

## Conclusion

This study investigated how general users conceptualize "robot knowledge" and how this framework differs from their understanding of human knowledge. Across three experiments, consistent evidence demonstrated that users enforce significantly more rigid and restrictive benchmarks when evaluating the epistemic states of a robot compared to those of a human. Specifically, observers maintain that a robot must directly perceive or verify information to achieve true knowledge, whereas they readily credit humans with the capacity for abstract inference and indirect knowing. These divergent expectations regarding knowledge acquisition hold meaningful implications for the future of human–robot interaction.

Experiment 1 established that even within structurally identical scenarios, users apply distinct evaluative criteria to knowledge acquisition depending on whether the subject is a human or a machine. Experiment 2 expanded this insight, demonstrating that users perceive robots as acquiring knowledge more slowly and strictly through direct observation, while humans are seen as capable of rapid knowledge acquisition via testimony. Experiment 3 linked these epistemic assumptions to moral evaluations: when robots functioned as accomplices within illicit scenarios, users ascribed less overall responsibility to them than to human accomplices. This shift is directly attributable to the user's belief that a robot's knowledge-acquisition capabilities are inherently limited.

Synthesizing these outcomes indicates that users do not view robots simply as unoptimized tools, but rather as fundamentally distinct "intelligent entities." This cognitive framing directly shapes how users evaluate a machine's autonomy, moral culpability, and operational reliability. As robots become increasingly integrated into social and

collaborative settings, these underlying beliefs will inevitably shape trust dynamics, responsibility attribution, and the broader human–robot partnership.

Ultimately, this research highlights the necessity of developing systems that actively manage user expectations regarding robotic knowledge within HRI (Human–Robot Interaction) and HRC (Human–Robot Collaboration) design. Engineers and designers may need to implement transparent communication interfaces that explicitly signal when and how a robot captures and processes information. Future research should prioritize investigating interaction design methodologies that can adjust these user assumptions, while exploring how such calibrations influence human cooperation, task delegation, and long-term trust in robotic systems.

## Acknowledgements

I give my deepest thanks to the participants in the experiment

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 21 Jul 2022

Revised: 23 Aug 2022

Accepted: 15 Sep 2022

Published online: 15 March 2023

## Rights and permissions

**Open Access** The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

## References

Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F. Social robots for education: a review. *Sci Robot*. 2018;3(21):eaat5954.  
<https://doi.org/10.1126/scirobotics.aat5954>.

Broadbent E, Stafford R, MacDonald B. Acceptance of healthcare robots for the older population: review and future directions. *Int J Soc Robot*. 2009;1(4):319–30.  
<https://doi.org/10.1007/s12369-009-0030-6>.

Chita-Tegmark M, Scheutz M. Assistive robots for the social management of health: a framework for robot design and human-robot interaction research. *Int J Soc Robot*. 2021;13(2):197-217.  
<https://doi.org/10.1007/s12369-020-00634-z>.

Hancock PA, Billings DR, Schaefer KE, Chen JYC, de Visser EJ, Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. *Hum Factors*. 2011;53(5):517-27.  
<https://doi.org/10.1177/0018720811417254>.

Yamagishi T, Kikuchi M, Kosugi M. Trust, gullibility, and social intelligence. *Asian J Soc Psychol*. 1999;2(1):145-61.

Robinette P, Howard AM, Wagner AR. Effect of robot performance on human-robot trust in time-critical situations. *IEEE Trans Human Mach Syst*. 2017;47(4):425-36.  
<https://doi.org/10.1109/THMS.2017.2676249>.

Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In: *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*; 2015. p.141-8.  
<https://doi.org/10.1145/2696454.2696497>.

Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors*. 2004;46(1):50-80.  
<https://doi.org/10.1518/hfes.46.1.50.30392>.

Eyssel F, Hegel F. (S)he's got the look: gender stereotyping of robots. *J Appl Soc Psychol*. 2012;42(9):2213-30.  
<https://doi.org/10.1111/j.1559-1816.2012.00937.x>.

Kahn PH Jr, Gary HE, Shen S. Children's social relationships with current and near-future robots. *Child Dev Perspect*. 2013;7(1):32-7.  
<https://doi.org/10.1111/cdep.12011>.

Waytz A, Heafner J, Epley N. The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. *J Exp Soc Psychol*. 2014;52:113-7.  
<https://doi.org/10.1016/j.jesp.2014.01.005>.

Matsui T, Koike A. Who is to blame? The appearance of virtual agents and the attribution of perceived responsibility. *Sensors (Basel)*. 2021;21(8):2646.  
<https://doi.org/10.3390/s21082646>.

Kawai Y, Miyake T, Park J, Shimaya J, Takahashi H, Asada M. Anthropomorphism-based causal and responsibility attributions to robots. *Sci Rep*. 2023;13(1):12234.  
<https://doi.org/10.1038/s41598-023-39435-5>.

Coeckelbergh M. Robot rights? Towards a social-relational justification of moral consideration. *Ethics Inf Technol*. 2010;12(3):209-21.  
<https://doi.org/10.1007/s10676-010-9235-5>.

Matthias A. The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf Technol*. 2004;6(3):175-83.  
<https://doi.org/10.1023/B:ETIN.0000048790.49568.7f>.

Taniguchi T, Nagai T, Nakamura T, Iwahashi N, Ogata T, Asoh H. Symbol emergence in robotics: a survey. *Adv Robot*. 2016;30(11-12):706-28.  
<https://doi.org/10.1080/01691864.2016.1164622>.

Gettier EL. Is justified true belief knowledge? *Analysis*. 1963;23(6):121-3.

Zagzebski L. The inescapability of Gettier problems. *Philos Q*. 1994;44(174):65-73.  
<https://doi.org/10.2307/2220058>.

Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspect Psychol Sci*. 2011;6(1):3-5.  
<https://doi.org/10.1177/1745691610393980>.

Goodman JK, Cryder CE, Cheema A. Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *J Behav Decis Mak*. 2013;26(3):213-24.  
<https://doi.org/10.1002/bdm.1753>.

Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. *Judgm Decis Mak*. 2010;5(5):411-9.

Malle BF, 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 10th ACM/IEEE International Conference on Human-Robot Interaction*; 2015. p.117-24.  
<https://doi.org/10.1145/2696454.2696471>.

Cushman F. Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*. 2008;108(2):353-80.  
<https://doi.org/10.1016/j.cognition.2008.03.006>.

Hart HLA. Negligence, mens rea, and criminal responsibility. *Oxf J Leg Stud*. 1968;8(3):322-49.

Malle BF, Guglielmo S, Monroe AE. A theory of blame. *Psychol Inq*. 2014;25(2):147-86.  
<https://doi.org/10.1080/1047840X.2014.877340>.

Robinette P, Li W, Allen R, Howard AM, Wagner AR. Overtrust of robots in emergency evacuation scenarios. In: *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*; 2016. p.101-8.  
<https://doi.org/10.1109/HRI.2016.7451740>.