How should an AI trust its human teammates? Exploring possible cues of artificial trust

Centeio Jorge, C.; Jonker, C.M.; Tielman, M.L.

DOI
10.1145/3635475

Publication date
2024

Document Version
Final published version

Published in
ACM Transactions on Interactive Intelligent Systems

How Should an AI Trust its Human Teammates? Exploring Possible Cues of Artificial Trust

CAROLINA CENTEIO JORGE, Delft University of Technology, The Netherlands
CATHOLIJN M. JONKER, Delft University of Technology and University of Leiden, The Netherlands
MYRTHE L. TIELMAN, Delft University of Technology, The Netherlands

In teams composed of humans, we use trust in others to make decisions, such as what to do next, who to help and who to ask for help. When a team member is artificial, they should also be able to assess whether a human teammate is trustworthy for a certain task. We see trustworthiness as the combination of (1) whether someone will do a task and (2) whether they can do it. With building beliefs in trustworthiness as an ultimate goal, we explore which internal factors (krypta) of the human may play a role (e.g., ability, benevolence, and integrity) in determining trustworthiness, according to existing literature. Furthermore, we investigate which observable metrics (manifesta) an agent may take into account as cues for the human teammate's krypta in an online 2D grid-world experiment (n = 54). Results suggest that cues of ability, benevolence and integrity influence trustworthiness. However, we observed that trustworthiness is mainly influenced by human's playing strategy and cost-benefit analysis, which deserves further investigation. This is a first step towards building informed beliefs of human trustworthiness in human-AI teamwork.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence;**

Additional Key Words and Phrases: Artificial trust, trustworthiness, teamwork, hybrid teams, human-AI teams, human-agent interaction

ACM Reference format:
Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. How Should an AI Trust its Human Teammates? Exploring Possible Cues of Artificial Trust. *ACM Trans. Interact. Intell. Syst.* 14, 1, Article 5 (January 2024), 26 pages.
https://doi.org/10.1145/3635475

5

This material is supported by Delft AI Initiative and by the TAILOR Connectivity Fund. Similarly, it is based upon work supported by the National Science Foundation (NWO) under Grant No. (1136993), and by the European Commission funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant 820437). The support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these institutions.
Authors' addresses: C. Centeio Jorge and M. L. Tielman, Delft University of Technology, Delft, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands; e-mails: {C.Jorge, M.L.Tielman}@tudelft.nl; C. M. Jonker, Delft University of Technology, Delft and University of Leiden, Niels Bohrweg 1, 2333 CA Leiden, Room number 126a; e-mail: C.M.Jonker@tudelft.nl.

2160-6455/2024/01-ART5
https://doi.org/10.1145/3635475

1 INTRODUCTION

Artificial agents are becoming more intelligent and able to execute relevant tasks for our daily lives, including in work environments, home assistance, battlefield and crisis response [39]. This holds for chat-based agents, intelligent virtual agents and even robots. These tasks should complement human's sensory and cognitive abilities. For example, an intelligent agent can quickly process large quantities of data, but it may require a human to make ethical decisions. In these cases, humans and intelligent agents should learn to cooperate, coordinate and collaborate with people, forming human-AI teams (also called human-agent, human-automation or human-machine teams). Humans make use of trust in each other (as well as trust in themselves) to make decisions and achieve effective teamwork, through communication and shared mental models [56]. For example, how I trust someone for a certain task, e.g., drive a car, will affect how I behave, e.g., I may accept a ride or suggest I drive instead. Similarly, we proposed in previous works [16] that AI teammates could make use of the human notion of trust to make decisions in human-AI teams, e.g., acting towards the team's goal and risk mitigation. For the AI to be able to form beliefs of artificial trust[1] in human teammates, we need to first study which characteristics make a human teammate trustworthy towards an AI teammate and how these characteristics can be perceived by the AI, as this is not present in literature to the best of authors' knowledge. To try to close this gap in literature, we suggest in Reference [17] which characteristics may form this trustworthiness and how these can be observed. This article extends this work by exploring how these metrics can actually be used in an online experiment involving humans teaming up with artificial agents.

Using notions of trust for artificial agents is in fact not new for **Multi-Agent System** (**MAS**), where artificial trust has been used among artificial agents for decision-making, see e.g., References [21, 55, 61]. However, we would like to similarly use artificial trust to enable AI teammates to delegate or decide how to rely on their human teammates, taking into account the team's goal and possible risks. In particular, artificial trust should help the AI teammate know (1) whether a human teammate could do a certain task and (2) whether they would actually do that task. Although artificial teammates are developed by humans and tailored to our needs, it is impossible to prepare them for all of their possible teammates and situations. Furthermore, people change with time and an artificial agent should be able to adapt throughout interactions. As such, artificial teammates should have the capacity to observe their human teammates and build beliefs regarding their trustworthiness, which will allow them to assist better. More specifically, the agent would be able to decide when to rely on someone and act accordingly, e.g., by helping the human or deciding on task allocation, mitigating the risks and ensuring the team's goal is reached [13]. We see reliance as the resulting behaviour of artificial trust evaluation, whereas artificial trust is a construct, i.e., a model composed of several aspects. Besides knowing the result of artificial trust evaluation (and, consequently, reliance), knowing which aspects constitute trust, i.e., by knowing why someone is or not trustworthy for a certain task, also contributes to better decisions and may improve the interaction between the human and the agent.

We approach artificial trust from a functional perspective, in which trust is a relational construct between the trustor x, the trustee y, about a defined (more or less specialized) task (τ), as in [22]. More concretely, artificial trust of x in y is x's belief about y's trustworthiness [16]. Literature so far explores how artificial agents can form beliefs regarding other artificial agent's trustworthiness, but not how they can form these beliefs regarding a human teammate. Thus, what makes a human trustworthy (towards AI) in a human-AI team setting, and how could an artificial agent observe it, given a specific task? We are presented with a large gap in the literature since:

[1]We use the term artificial trust as in Reference [6] to refer to AI's computation of trust in other agents or humans. We recognize that an agent's computational assessment of someone's trust differs from the human phenomenon of trust.

(1) There is no theory of what human's trustworthiness towards an artificial teammate is (i.e., what are the aspects of the construct) from social sciences' perspective.
(2) There is, consequently, little to no research regarding how these aspects that may compose this trustworthiness manifest (i.e., behavioural cues).
(3) No research has shown how this observable behaviour could be used for the formation of artificial beliefs regarding human's trustworthiness in a specific context and how these can be used.

As such, in this article, we take a step towards answering these questions by exploring how manifested (i.e., observable) behaviour could be used to establish different aspects of a human's trustworthiness, and how those relate to self-reported trustworthiness and overall success metrics. We depart from theories both in social sciences and multi-agent systems and investigate them through an experiment where 54 participants collaborated with simple artificial agents by collecting products from a supermarket in a 2D grid world (inspired by search and retrieve task, such as Blocks World for Teams [30] experiments). During the experiment, we collected logs of participant's behaviour as agent observations, and self-reported measures regarding participant's trustworthiness and goals in the experiment.

This work contributes by:

(1) Theoretically exploring through a multidisciplinary perspective:
  (a) How an artificial agent could break down a trustworthiness belief into different aspects (partly presented in conceptual model from Reference [17]);
  (b) How such an agent could form beliefs of these aspects regarding human's trustworthiness through observations;
(2) Presenting an experiment design which empirically explores how these aspects could be observed and how they relate to each-other given a specific task and scenario;
(3) Analysing through Bayesian statistics how these observations relate with overall success measures and human's self-reported trustworthiness;
(4) Reporting important transversal methodological challenges that may affect the study of such question;
(5) Relating these findings with human strategy to determine the next steps in allowing an artificial agent to form trust in human teammates.

The rest of this article is organized as follows: in Section 2, we explore the literature and concepts behind our model, then explain the experiment design in Section 3, and the results in Section 4. We then discuss the results in Section 5, summarizing the model in Section 3, and finally conclude in Section 6.

2 ASPECTS OF ARTIFICIAL TRUST

Most research on human–machine interaction has focused on how humans trust artificial agents, see e.g., References [24, 27, 33, 38, 45, 47, 64] and not vice versa. However, there is some work in this direction, for instance, how an artificial agent can detect whether a human is being trustworthy, based on episodic memory [63], i.e., based on how many times the human was reliable in the past, and on social cues from video of a human interacting with a robot [60]. Also, Reference [6] has proposed a model to predict how much humans can be trusted to execute a task, in human–robot teams, focusing only on a human's capabilities. None of these works has tried to deconstruct human trustworthiness, however, but rather looked at it as a simple metric, and mainly focusing on performance. Instead, we propose that we should take several dimensions into account when determining trustworthiness. By learning the mental model of the human teammates, we believe

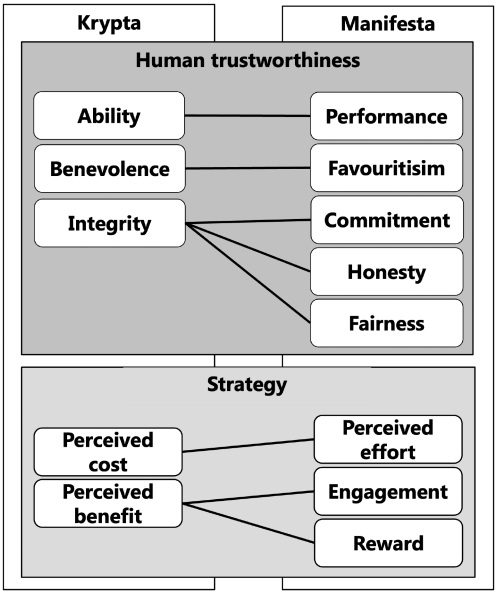

Fig. 1. Krypta and manifesta of human trustworthiness in human-AI teams, part of conceptual model previously presented in Reference [17].

the agent is better equipped to make informed choices, mitigating risks and eliciting appropriate trust while still assessing trustworthiness. In this section, we present the theory behind our proposed mental model of human trustworthiness in human-AI teams (in Figure 1). This theory is based on existing concepts within the literature.

We start by exploring how artificial agents could use artificial trust, from a computational perspective. Reference [21] proposes that artificial trust can be deconstructed in two beliefs regarding trustee's trustworthiness, i.e., *competence* belief, and *willingness* belief. The competence belief reflects an evaluation of the trustee's abilities, meaning that the trustee can produce the expected results (i.e., can perform an action as expected). On the other hand, the willingness belief translates to whether the trustor believes the trustee will do the task (independently of competence belief). These beliefs may be affected by *external factors* like opportunities and interferences [21], which can be part of activity context and process [29, 37].

As we do not know how humans manifest willingness and competence directly (willingness is particularly difficult), we start by exploring which internal features makes the human more or less trustworthy, and how these could be observed through behaviour. We follow [22], who propose that trust beliefs are formed from the observable behaviour of an agent, the *manifesta*, which are signals that indicate certain internal features, the *krypta* (inspired by Reference [8]). In Section 2.1.1, we establish which krypta to use based on human–human trust models, and in Section 2.1.2, we propose how to observe these in a human-AI teamwork scenario.

We also explore which factors are important in human's strategy in Section 2.2. Factors such as preference and perceived risk are often mentioned as elements that affect decision-making in general [29, 42]. We claim that some of these factors form human *strategy*. Strategy is mainly related to the goal, the task, and the consequence of taking the task. It also plays a role in the decision-making of the trustee, determining whether a task will be performed, thereby affecting the trustee's trustworthiness.

In the following sections, we explore the relationships between manifesta, krypta and strategy.

2.1 Human Trustworthiness

2.1.1 Krypta. Krypta is the set of internal features of an agent [8] that make them more or less trustworthy. When transferring these notions to humans, we base our human krypta on the *ABI* model of trust [42], which has been widely used to study trustworthiness in organizational psychology. Although we do not know if this is the adequate krypta for human trustworthiness in human-AI teams, this is the closest we find in literature. *ABI* says that human trustworthiness depends on their internal features of ability, benevolence and integrity. The authors define trust as *the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party* (p. 712). In this model of trust, trustworthiness is defined as *the extent to which an actor has the ability to execute relevant tasks, demonstrates integrity, and is benevolent towards fellow team members* [64] (p. 461). Furthermore, these are the definitions of ability, benevolence and integrity that can be found in Reference [42]:

— *Ability is that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain.* (p. 717)
— *Benevolence is the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive. Benevolence suggests that the trustee has some specific attachment to the trustor.* (p. 718)
— *The relationship between integrity and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable.* (p. 719)

Compared with the definitions of competence and willingness from Reference [21], we can associate ability with competence, and all three (ability, benevolence and integrity) with willingness. The next and final step of building our mental model is to explore ways of building a manifesta, i.e., behaviours which can give us cues to the krypta, from the literature.

2.1.2 Manifesta. We looked into literature to find possible ways of observing ability, benevolence and integrity, so that we can have cues of the human krypta, and finally form beliefs regarding human's trustworthiness (i.e., willingness and competence).

Ability in the context of human teams can be observed in how successfully a task is performed (e.g., based on time or score of some kind), how much effort was put to do a task well, by continuously working thoroughly and accurately, and also in how appropriately the tools (such as technology) were used [12].

Benevolence can take its time to meaningfully develop [42], since it is connected to the relationship between the trustor and trustee. This makes the process of observing it in first-time interactions hard. In multi-agent systems, a benevolent agent is the agent that accepts the requests of other agents, i.e., the one that voluntarily helps another agent, without this serving or harming its own goal [40, 46]. We can then observe it through task support, i.e., when a teammate helps another by helping or completing a task [12]. Benevolence is then intertwined with the personal relationship between the trustor and the trustee, i.e., it has to do with the specific altruistic attitude of the trustee regarding a certain trustor.

Integrity, finally, is by definition related to values and moral principles. These principles can be such as honesty, truthfulness, sincerity, fairness, and ability to keep commitments (i.e., reliability, dependability) [44, 51]. As such, we can observe it through credible communications, a strong sense of justice, consistency of word and action, and availability [2, 12, 42]. It differs from benevolence, since it is related to general principles and values of the trustee, rather than trustee's attitude towards the particular trustor. However, depending on the literature, there are traits that are sometimes associated with benevolence and other times with integrity, which is the case of

commitment and availability, for example. In fact, Reference [54] presented a schema in which both availability and commitment are considered to be an antecedent of benevolence.

2.2 Strategy

2.2.1 Krypta. During our pilot studies, we could observe that participants might be following a *strategy*, i.e., to select their perceived advantageous alternatives from the beginning and persist on these lines of options [11]. However, what is advantageous for one participant might not be for another. In fact, human decision-making is influenced by explicitness of positive and negative consequences as well as the directness of probabilities for reward and punishment [11]. Although this is not the main focus of this study, we believe it should still be addressed. For this reason, we can see in Figure 1 a block for *strategy*, where *perceived cost* and *perceived benefit* are addressed as the main krypta factors, not directly observable. Perceived cost-benefit is affected by several factors, including goals, motivation, engagement, perceived risk, perceived effort, difficulty, time, utility, and overall cognitive characteristics [35, 50, 57]. Overall, what is effort and how a certain effort is rewarding to us depends on our characteristics (krypta) [26] (e.g., a person with good photographic memory may find it effortless to collect a new product).

2.2.2 Manifesta. How an agent can observe the perceived cost and perceived benefit is still an open question, as well as the relationship with the three trustworthiness dimensions. We do speculate, however, that the agent might be able to calculate perceived effort, engagement and reward, through observation of repeated human behaviour (see e.g., References [20, 32, 49]. How the strategy can be observed will not be the focus of the design of the experiment, but it will be further explored in the discussion (Section 5).

2.3 Summary

In this section, we explored the theory that indicates how we can measure ability, benevolence and integrity. However, we still need to investigate how these can be in practice applied to human-AI teamwork and how they should manifest. In particular, we aim at filling the gap in the literature by exploring how to observe trustworthiness's dimensions from humans, in human-AI teamwork. We hypothesise that an agent can build trustworthiness beliefs of a human teammate's ability, benevolence and integrity (the krypta) based on observations of human behaviour (the manifesta). Because it is challenging to compare our observations to a ground truth (if we could understand trustworthiness perfectly, it would not be a challenge for an agent either), we cannot prove our hypothesis. However, we can and will explore how our observations relate to self-reported trustworthiness and general metrics of success (which are direct consequences of trustworthiness). We do not claim that people have a perfect perception of their own trustworthiness, but rather are interested in exploring the relationships between self-reported and observed behaviour. Besides this main focus of our article, we want to also investigate which other factors, part of human's strategy, might influence decision-making in this setting.

3 METHOD

We have conducted an experiment to explore how an agent can form beliefs regarding its human teammate's trustworthiness. The goal of this experiment is to explore how we can observe behaviour that is associated to ability, benevolence, and integrity. This experiment was done online, where participants accessed through their browser from their homes, while on a call with the experimenter. We collected data through logs regarding the human's observable behaviour (i.e., choices, performance, etc) that we relate to ability, benevolence and integrity, as well as human's self-reported (subjective) metrics regarding their own ability, benevolence, and integrity during the experiment.

3.1 Participants

This research received ethical approval from the TU Delft HREC, nr 1672. Fifty-four participants were recruited using the authors' personal networks and, in some cases, participants recruited further participants. The most frequent age group was 25–34 (42 participants) and ages were between 18 and 54 years old. Two-thirds of the participants identified themselves as Male and the rest as Female. The participants' cultural background was mostly European (43) and their experience with computer games ranged from low (11), and average (24) to advanced (19).

We first used four of the participants for the pilot. After the pilot, we added one final question regarding the strategy (see in Section 4.4) to the experiment, as explained in Section 2.2. For this reason, we used the data collected during the pilot in the analysis, except in the last question.

3.2 Environment

To observe ability, benevolence and integrity in human-AI teams, we needed a task which was accessible to all participants, but could differentiate them along the three dimensions. As such, the task was easy but (1) required some memory and keyboard ability as additional competences (so we could observe ability), (2) presented two different agents asking for collaboration (so that we could see benevolence), and (3) gave the participants the freedom to lie, give up and be fair (so we could see integrity). In this experiment, artificial agents asked their human teammates to collect products in a 2D grid world supermarket (inspired by the booming of click&collect shopping during pandemic) developed using Matrx package[2][3]. The environment consisted of the supermarket (Figure 2 where the participant interacted with the products and a chat, in which the participants could interact with the agent.

Participants (marked as a yellow smiley) were asked to imagine themselves as workers, with the role of collector, in the supermarket. (Imaginary) Customers ordered from the supermarket online. These orders were processed by artificial Agent X and Agent Z, who were the participant's teammates (marked as yellow smiley with glasses, standing next to a basket and a letter "X" or "Z", respectively). During the experiment, the agents showcased the product that needed to be collected in the light blue area below them and announced it in the chat. These products were disposed in the several aisles of the supermarket and may not be visible to the participant from the distance, depending on participant's virtual capabilities (based on group characteristics, as explained in Section 3.3). The participant could access the chat and communicate with the agents through buttons "Help X", "Help Z", "Collected", and "Give up". The stochasticity present in this experiment is limited to the products that appeared in the blue areas. In order to keep control of the environment and different conditions, the agents present in this experiment were not intelligent agents, i.e., they did not have autonomy nor learned actively. However, the participant does not know the level of autonomy of the agent, so we do not think this affects how the participant perceived the AI.

3.2.1 Task. The participant's job was to help the agents collect as many products as possible, during 10 minutes. Participants could check the products presented by the two agents and choose which one to help (only one at a time), by pressing the button "Help [agent id]" (i.e., to help Agent X, participant should press "Help X"). The other (i.e., Agent Z in this case) took it as a rejection and presented another product (it counted as if the product was collected by another agent, so that the participant could choose freely who they wanted to help). After committing to helping an agent, the participant was expected to search for the product (participants moved with keys), bring it

[2]https://www.matrx-software.com/

[3]The code and raw data can be found in https://github.com/centeio/click-collect and https://doi.org/10.4121/21982991.v1.

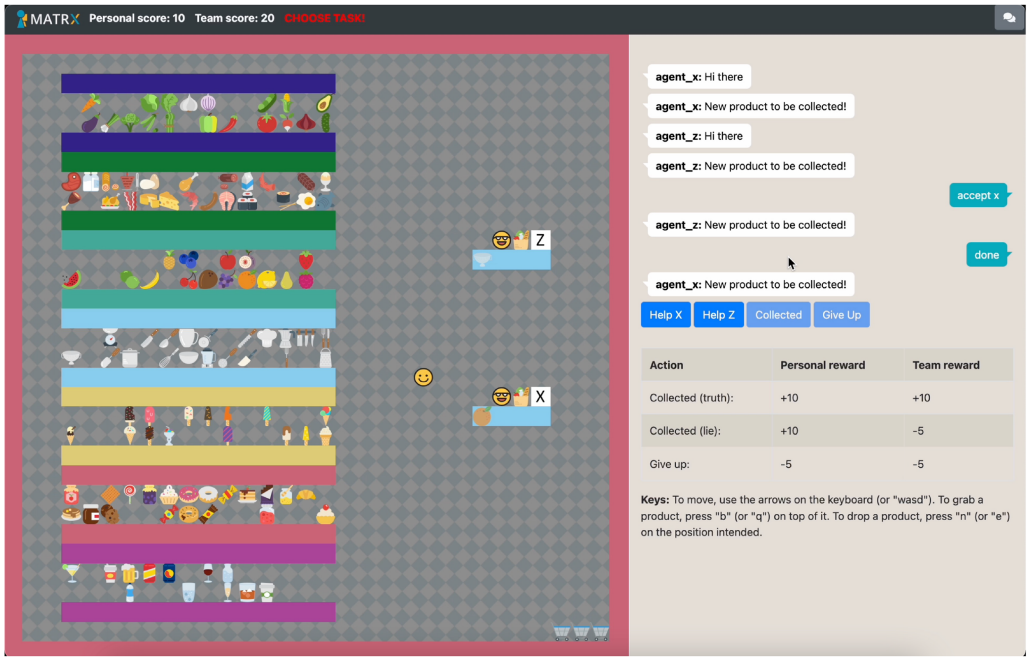

Fig. 2. Experiment environment, a 2D grid world supermarket. On the left side of the screen, the participant (smiley face without glasses) had to help the agents (the smiley faces with glasses) collecting products. They could communicate using the buttons on the chat side of the page (right side). Scores appear in the top left corner.

to the agent and press the button "Collected" (counted as a *success*). "Collected" could be pressed even if the product was not actually collected, allowing participants to lie to the agent (counted as a *lie*). They could also give up on a task by pressing the button "Give up" (counted as *give up*). Agents presented new products every time the participant gave up on, collected or rejected their product.

In this setting, it was only possible to succeed at the task, without, otherwise being counted either as a *lie* or as a *give-up*. We made this choice because we wanted to know the reason why a task was unsuccessful. According to the pilot, the task was quite accessible, and the best someone did, the more tasks that person would complete and more quickly. They would either purposely deliver the wrong product and say it was collected (in which case it is a lie) or decide to give up. As such, we counted for the number of successfully completed tasks as an indicator of participant's success.

Furthermore, there were two scores in the game for the participant, the *personal score* and the *team score*. Rewards follow Table 1. We separated the score into team and personal to accommodate the idea that sometimes personal and team goal may collide. In this case, when a person lied, they could still get the *success* reward for personal score, but it harmed the team, reducing the team's score by 5 points. Different subjects may care differently about team or personal score.

3.3 Conditions

We divided the participants in four different groups. Our main objective was not to know how our manipulations affected trustworthiness directly, but rather to provide some possibility of variance among ability, benevolence and integrity (as we believed there was the risk that the environment

Table 1. Reward System

Action	Success	Lie	Give up
Personal	10	10	−5
Team	10	−5	−5

was not complex enough to show this variation without manipulation). As these groups were not thought to be compared, this was **not** a comparative experiment. The groups were:

(1) *Gn*: This group was presented to the environment without adding anything to the narrative. They could only see the products when standing from a certain distance.
(2) *Ga*: This group was given better virtual abilities, i.e., this was the only group that could see all products in the supermarket, regardless of distance. The objective was to diversify ability. We expect participants of this group to be able to complete the tasks faster, by finding the products first, for example.
(3) *Gb*: This group was asked to imagine they had a close relationship of colleague/friends with Agent X, whereas Agent Z was new to the supermarket (based on Reference [1]). Furthermore, based on Reference [48], we gave Agent X characteristics that motivated mindless behaviour, by giving the Agent X's and the human's avatar the same colour (green), instead of yellow (as Agent Z's), and by giving Agent X a friendlier way of talking. The objective was to diversify benevolence. We expect participants of this group to lean towards helping one agent more over the other, showing higher benevolence towards one specific agent.
(4) *Gi*: This group was asked to imagine they were in a temporary job and that they would earn money proportionally to their personal score. We expect participants of this group to prioritize their personal motives, e.g., wanting to do the highest amount of tasks in the shortest amount of time possible. This can lead to participants giving up more often, i.e., if they realize they cannot complete a certain task quickly, they may want to go for the next one. Also, these participants may lie to the agent more often, by saying that a task is completed when it is actually not, as they would get the same reward, which would then convert in the money they would hypothetically get. The main objective was to diversify the priority of principles among participants, thereby diversifying integrity.

3.4 Procedure

Foremost, each participant accessed to a previously shared online meeting room. The participant was then asked first fill in the consent form online, and proceed to the demographic questions regarding their gender, age, cultural background, and expertise in computer games. Then, they were randomly assigned to one of the four conditions, and the researcher explained the task. They could then try it out in a trial environment, where the researcher made sure they understood the task and how to navigate in the environment. Finally, the participant played the task for 10 min and after that answered the questions regarding their own trustworthiness.

3.5 Subjective Measures

3.5.1 Self-Evaluation of Trustworthiness. We adapted the validated questionnaire of Reference [1] (also based on References [2] and [41]) regarding trust in (military) teams to perceived own trustworthiness, as to have subjective measures of the participant's self-estimation of ability (in the original questionnaire as capability), benevolence and integrity. The questions regarded (1) capability/ability, relating this dimension to self-perception of capability, knowledge, qualification and communication and faith in self's abilities; (2) benevolence, where participants were inquired about their attitude specifically towards each of the agents, in terms of having

agent's best interests in mind, looking out for the agent, and working/wanting to protect the agent; (3) integrity, where participants were self-evaluated regarding fairness, honour, honouring their word, keeping promises and telling the truth. We referred to the questions regarding ability as *QA* (which go from *QA1* to *QA5* and have a *QA Mean* of the five questions), benevolence towards Agent X as *QB* and integrity *QI* in a similar fashion. The sum of all measures per participant (subjective trustworthiness towards X) was *STW*. In summary:

— *QA Mean*: average of QA1, QA2, QA3, QA4, and QA5.
— *QB Mean*: average of QB1, QB2, QB3, QB4, and QB5.
— *QI Mean*: average of QI1, QI2, QI3, QI4, and QI5.

3.5.2 Strategy. Finally, at the end of the questionnaire, participants were asked what their goals were during the experiment. This question was added in order to explore what might be behind a participant's strategy choice. It was a multiple choice question (allowed to tick more than one box), in which the options were elaborated mainly based on different perceptions of cost and benefit of the world, including perceived effort and value attributed to the scores. As the task was part of a teamwork scenario, and this was highlighted in certain conditions, we also added options regarding their teamwork, where the strategy would mainly focus on helping the agents. The options were: "Collect as many products as possible", "Collect products as fast as possible", "Maximize personal score", "Maximize team score", "Collect the easiest products (based on icon)", "Collect easiest products (based on distance)", "Collect according to the chat messages", "Helping both agents equally", "Helping specifically agent X", "Helping specifically agent Z", "I do not know", and "Other". When choosing "Other", they could write a goal in their own words. This question was mainly exploratory.

3.6 Agent Observations

Based on the subjective measures in Section 3.5 and literature in Section 2, we chose the human teammate's manifesta (as presented in Figure 1). These measures should translate the agent's observations into the concepts related to the definitions of ability, benevolence and integrity, in a similar way to how they were represented in the questionnaire. In this experiment, we logged the main actions of the participants, with a timestamp number of moves since the start of the experiment, and a Manhattan distance between the product position in the supermarket and an average participant's starting position. The events being logged were:

— A newly presented task by an agent, i.e., when an agent asked for a new product to be collected by the participant. This happened every time (1) the participant declined their task (by accepting the task of the other agent), (2) concluded successfully the task or (3) concluded unsuccessfully the task.
— The participant accepted one agent's task, by pressing "Help".
— The participant concluded the task, whether it was by pressing "Collected" (which registers whether this was successful, which counted as a success, or unsuccessful, which counted as a "Lie") or "Give up".

With these logs, we calculated the number of presented tasks, accepted tasks, successful tasks, lies, give-ups per agent for each participant. We also calculated average times and moves. Using these, we computed the agent observations of ability, benevolence, and integrity. Although some of the measures may relate with more than one of the three definitions of ability, benevolence, and integrity, we related them with the one that was closest to definition and questionnaire questions.

3.6.1 Ability. The observable metrics related to ability were:

— *Time/Task*: time spent per successful task. This was an indicative of higher performance when lower, meaning that someone needs less time to complete a task.
— *Moves/Task*: steps needed to successfully complete a task. Just like *Time*, it was also an indication of higher performance when lower, as the subject needed fewer moves to find the product and collect it (and return it). However, certain tasks required more moves than others (since the products were randomly selected in the grid). Thus, we calculated this metric as

$$\frac{Moves/Task}{Mean\ task\ difficulty}, \tag{1}$$

where *Mean task difficulty* is the shortest path to the required product).
— *Moves/Time*, i.e.,

$$\frac{Moves/Task}{Time}. \tag{2}$$

This metric should indicate better performance when higher, meaning the subject moved fast.

3.6.2 Benevolence: As benevolence relates to whether the trustee wants good to the trustor, we have used as metric a *Favouritism* factor, which was the ratio of number of successful tasks per agent, i.e.,

$$\frac{\#\ successful\ tasks\ for\ Agent\ X}{\#\ successful\ tasks\ for\ both\ Agents}. \tag{3}$$

This indicated a participant helped more (favoured) one of the agents.

3.6.3 Integrity: For integrity, we looked at it from different perspectives, and combined the factors believed to affect it. We computed:

— *Honesty* factor: number of lies over total tasks: i.e.,

$$\frac{\#\ Lies\ for\ both\ Agents}{\#\ Accepted\ Tasks\ for\ both\ Agents}. \tag{4}$$

— *Commitment* factor: number of given up tasks, i.e.,

$$\frac{\#\ Give\text{-}ups\ for\ both\ Agents}{\#\ Accepted\ Tasks\ for\ both\ Agents} \tag{5}$$

Although the number of give-ups can also indicate a lack of ability, we chose to associate it with integrity since, as explained before, we considered the task feasible (and all participants tried it through a tutorial first). As such, when a person decided to give-up, it showed more about persistence and "keeping promises", which are traits of integrity according to the questionnaire used.
— *Fairness 1*: According to the dictionaries of Cambridge and Merriam-Webster, fairness can be defined as treating people equally, impartially, free from self-interest, prejudice or favouritism. As such, for *Fairness 1*, we calculated the absolute difference between the lies given to each agent, i.e.,

$$abs\left(\frac{\#\ Lies\ for\ Agent\ X}{\#\ Accepted\ Tasks\ for\ Agent\ X} - \frac{\#\ Lies\ for\ Agent\ Z}{\#\ Accepted\ Tasks\ for\ Agent\ Z}\right) \tag{6}$$

This fairness factor aimed at reflecting the fairness w.r.t. honesty. Although the difference of lies between agents could be interpreted as a sign of higher benevolence (towards the agent the participant lied the least), we considered this to be a signal of integrity since the

participant was harming one more than the other (which is unfair). The difference between this metric and Favouritism is that Favouritism is something positive, such as helping out a friend (not necessarily with the intention of harming the other).

— *Fairness 2*. Similarly, this factor was absolute difference between the give-ups towards each agent, i.e.,

$$abs\left(\frac{\#\ Give\text{-}ups\ for\ Agent\ X}{\#\ Accepted\ Tasks\ for\ Agent\ X} - \frac{\#\ Give\text{-}ups\ for\ Agent\ Z}{\#\ Accepted\ Tasks\ for\ Agent\ Z}\right) \tag{7}$$

This fairness factor aimed at reflecting the fairness w.r.t. commitment.

3.6.4 Trustworthiness: Although we frame trustworthiness as a combination of all the above, we also computed the direct consequences of it through success metrics. This was mainly useful to speculate how other metrics impacted overall trustworthiness in each situation. We see the consequences of trustworthiness in terms of the success of the tasks, which is usually the main goal in teamwork situations. In case of other goals, other metrics for trustworthiness may apply. In particular, we divide success as a consequence of trustworthiness in two ways:

— *TW abs*: This is the absolute consequence of trustworthiness, and it was calculated by *# Successful tasks to Agent X*. We called it absolute because it is the raw number of successes during the 10 minutes of experiment. This has to do with how well a participant can do a task, assuming that the more they successfully complete in 10 minutes, the fastest they will do it (which, in this task, is the "how well" indicator).

— *TW rel*: The relative consequence of trustworthiness was calculated as the ratio of presented tasks that were successful, i.e.,

$$\frac{\#\ Successful\ tasks\ to\ Agent\ X}{\#\ Presented\ tasks\ by\ Agent\ X}. \tag{8}$$

This was relative as it could be seen as the probability of one succeeding at a task when asked. Thus, this suggests whether the participant would do a task when asked.

4 RESULTS

In this section, we report how the observations (manifesta) relate with each other, and how these relate to participants' self reports (of ability, benevolence and integrity). As mentioned before, the purpose of separating the participants per condition was to create variation in the participants' manifestation of ability, benevolence and integrity, though manipulation of narrative or environment, but not of the task. The conditions were not created so that we could evaluate each condition against a control group, necessarily, since we can group them and see the relationship among variables. Still, we compared the metrics among the conditions to see the effect of our manipulation. We used R 4.2.2, with the packages First Aid 0.1 [7] for Bayesian *t*-tests and correlations.

Bayesian methods have become more popular when analysing behaviour data, see e.g., References [3, 4, 9, 23, 52]. These have been found as an alternative to the more popular Frequentist tests. Frequentist approaches usually try to prove a null-hypothesis through frequentist statistical tests (which many times require certain assumptions from the data) and a *p-value*. Only with a low enough *p-value* we can say something about our hypothesis. This can be very hard to obtain with low quantity of behavioural data, which we usually get when doing research in human-computer/human-robot interaction. Furthermore, we cannot really say how likely this is to be a good hypothesis, only that it is (or not) *statistically significant*. Bayesian tests, on the other hand, present probabilities, e.g., how likely it is that there is actually a difference between samples

(instead of a yes/no). For these and other reasons, we chose to use Bayesian *t*-test and Pearson correlation for our data. In this article, specifically, we do not try to prove one hypothesis. Instead, we explore how the subjective measures and the agent observations relate and whether there was any difference among the conditions.

For Bayesian tests, both *t*-test in Section 4.2 and Pearson correlation in Section 4.3, several possible normal distributions that may fit to each of the metrics of our data are computed. This means that each metric will have a distribution of credible means and standard deviations. The test formula is then used for each of the credible combinations of means and standard deviations. Thus, the test results will also have an average value. We report all these by their 95% **High Density Interval** (**HDI**). Finally, we will evaluate the results of the tests by the probability of this average being positive (meaning there is or not a difference) and interpret it according to Reference [34]. The closer to 0 or 1 the probability is, the more significant it is (depending on whether it is negative or positive difference, respectively). When this probability is around 0.5, it means that this average is around 0, which tells us that there is probably no difference between the two groups for that formula. More on how Bayesian tests work can be found in References [34, 43].

4.1 Subjective Measures

The original questionnaire from Reference [1] is a validated one, where the authors ran both Exploratory Factor Analysis and Confirmatory Factor Analysis. However, since we adapted the tool, we also ran a Cronbach alpha on our results. We found good Cronbach's alphas [19] for ability questions ($\alpha = 0.89$) and integrity questions ($\alpha = 0.84$), and excellent ones for benevolence questions ($\alpha = 0.93$).

4.2 Differences Between Conditions

In this subsection, we look at the differences of the means of subjective measures and objective (observed) metrics of ability, benevolence, integrity and overall trustworthiness between conditions. We compare each group (Ga, Gb, Gi) with Gn. In particular, in Table 2 we compared Gb and Gn's metrics related to benevolence, both subjective (questionnaire benevolence-related items *QB* from 1 to 5 and the mean) and observed (*Favouritism*). Similarly, Table 3 compares groups Ga and Gn in terms of ability metrics from questionnaire (QA1 to QA5 and mean) and observations related to time and moves. Finally, Table 4 compares groups Gi and Gn in terms of integrity metrics, both subjective (questionnaire items *QI* 1 to 5 and mean) and objective (*Honesty*, *Commitment*, *Fairness 1* and *2*).

In each of these tables, the first two columns show the average of the metrics for each of the groups, with the limits of 95% HDI within brackets. The difference between means is showed in Diff Means column, also with the limits of 95% highest density interval within brackets. The SD columns show standard deviations in a similar fashion. Finally, the column % presents the probability of *Diff Means*> 0 and Evaluation column interprets the % column according to Reference [18]. This evaluation present the risk of betting on such correlation, which can translate into how probable a correlation is.

4.3 Correlations

After exploring the differences between conditions, we treated the dataset as one. In this subsection, we show the results of Bayesian Pearson correlation between metrics. Table 5 presents the correlations between subjective and observed metrics. Furthermore, Table 6 presents the correlations between observed consequences of trustworthiness metrics and observed metrics of ability, benevolence and integrity.

Table 2. Bayesian *T*-Test Between Benevolence (Gb) and Normal (Gn) Groups, Presenting Group's Possible Distributions' Means and Standard Deviations, the Difference Between the Means, the Probability of this Difference Being Positive and the Evaluation of this Probability, According to Reference [18]

Metric	Gb Mean	Gn Mean	Diff Means	Gb SD	Gn SD	%	Evaluation
QB1	4.8 [3.7, 5.9]	4 [2.5, 5.5]	0.8 [−1, 2.7]	1.8 [1.1, 2.9]	2.2 [1.2, 3.6]	0.8210	Casual bet
QB2	4.3 [3.4, 5.2]	3.4 [1.9, 5.1]	0.8 [−1, 2.6]	1.4 [0.8, 2.3]	2.4 [1.4, 3.9]	0.8282	Casual bet
QB3	4.4 [3.8, 5.1]	3.4 [1.8, 4.9]	1.1 [−0.6, 2.8]	1.1 [0.6, 1.7]	2.3 [1.3, 3.8]	0.9045	Promising but risky bet
QB4	4.8 [4.1, 5.5]	3.7 [2.4, 5]	1.1 [−0.3, 2.6]	1.2 [0.7, 1.8]	1.9 [1.1, 3]	0.9446	Promising but risky bet
QB5	4.7 [4, 5.4]	3.7 [2.4, 4.9]	1.1 [−0.4, 2.5]	1.1 [0.7, 1.8]	1.9 [1.1, 3]	0.9347	Promising but risky bet
QB Mean	4.6 [3.9, 5.4]	3.6 [2.3, 5]	1 [−0.6, 2.5]	1.2 [0.7, 1.9]	2 [1.1, 3.2]	0.8999	Casual bet
Favouritism	0.6 [0.5, 0.7]	0.6 [0.4, 0.7]	0.1 [-0.1, 0.3]	0.2 [0.1, 0.3]	0.2 [0.1, 0.4]	0.7661	Casual bet

As explained in Section 3.5, *QB* 1 to 5 correspond to each question regarding benevolence towards Agent X and *QB Mean* is their average. Favouritism is the observed metric related to benevolence.

Table 3. Bayesian *T*-Test Between Ability (Ga) and Normal (Gn) Groups, Presenting Group's Possible Distributions' Means and Standard Deviations, the Difference Between the Means, the Probability of this Difference Being Positive and the Evaluation of this Probability, According to Reference [18]

Metric	Ga Mean	Gn Mean	Diff Means	Ga SD	Gn SD	%	Evaluation
QA1	6.2 [5.5, 6.9]	5.5 [4.5, 6.4]	0.7 [−0.5, 1.9]	1 [0.5, 1.7]	1.4 [0.8, 2.3]	0.9019	Promising but risky bet
QA2	5.7 [4.9, 6.5]	5.2 [4.3, 6.1]	0.6 [−0.6, 1.8]	1.2 [0.7, 1.9]	1.3 [0.8, 2.2]	0.8355	Casual bet
QA3	5.4 [4.6, 6.3]	5.5 [4.6, 6.4]	0 [−1.3, 1.2]	1.3 [0.8, 2.1]	1.3 [0.7, 2.1]	0.4742	Not worth bet against
QA4	5.6 [4.7, 6.6]	5.4 [4.3, 6.5]	0.2 [−1.2, 1.7]	1.4 [0.8, 2.3]	1.6 [0.9, 2.7]	0.6309	Not worth bet on
QA5	5.7 [4.6, 6.8]	5.9 [5.2, 6.5]	−0.2 [−1.5, 1]	1.6 [0.9, 2.6]	1 [0.6, 1.6]	0.3633	Not worth bet against
QA Mean	5.7 [5, 6.5]	5.5 [4.6, 6.3]	0.2 [−0.9, 1.3]	1.1 [0.6, 1.7]	1.3 [0.7, 2.1]	0.6783	Not worth bet on
Time/Task	50.8 [27.1, 77.2]	70.6 [32.3, 110.1]	−19.4 [−65.9, 24.5]	35.6 [11.6, 63.5]	55.4 [27.6, 95.2]	0.1819	Casual bet against
Moves/Task	49.7 [44.5, 55.2]	59.5 [38.6, 81.4]	−9.8 [−31.6, 12.4]	8 [4.3, 13.5]	31.1 [17.5, 51.1]	0.1753	Casual bet against
Moves/Time	1.2 [0.8, 1.5]	1.2 [0.6, 1.6]	0 [−0.6, 0.6]	0.5 [0.3, 0.8]	0.7 [0.4, 1.3]	0.5049	Not worth bet on

As explained in Section 3.5, *QA* 1 to 5 correspond to each question regarding ability and *QA Mean* is their average. Time/Task, Moves/Task and Moves/Time are the observed metrics related to ability.

Table 4. Bayesian *T*-Test Between Integrity (Gi) and Normal (Gn) Groups, Presenting Group's Possible Distributions' Means and Standard Deviations, the Difference Between the Means, the Probability of this Difference Being Positive and the Evaluation of this Probability, According to Reference [18]

Metric	Gi Mean	Gn Mean	Diff Means	Gi SD	Gn SD	%	Evaluation
QI1	5.9 [5.4, 6.4]	5.3 [4.1, 6.5]	0.6 [−0.7, 1.9]	1 [0.7, 1.4]	1.8 [1.1, 2.9]	0.8286	Casual bet
QI2	6.1 [5.5, 6.7]	6.1 [5.3, 7]	0 [−1.1, 1]	1.2 [0.7, 1.7]	1.3 [0.7, 2.2]	0.4792	Not worth bet against
QI3	6 [5.5, 6.5]	6.1 [5.2, 7]	−0.1 [−1.2, 1]	1.1 [0.7, 1.5]	1.4 [0.8, 2.4]	0.4083	Not worth bet against
QI4	6.3 [5.7, 6.9]	7 [7, 7]	−0.7 [−1.3, -0.1]	0.8 [0.4, 1.4]	0 [0, 0]	0.0254	Promising but risky bet against
QI5	6.7 [6.1, 7.1]	7 [7, 7]	−0.3 [−0.9, 0.1]	0.6 [0, 1.1]	0 [0, 0]	0.0844	Promising but risky bet against
QI Mean	6 [5.6, 6.5]	6.1 [5.4, 6.9]	−0.1 [−1, 0.7]	0.9 [0.6, 1.2]	1.1 [0.6, 1.8]	0.4092	Not worth bet against
Honesty	1 [0.9, 1]	1 [1, 1]	0 [−0.1, 0]	0.1 [0, 0.1]	0 [0, 0]	0.0972	Promising but risky bet against
Commitment	1 [1, 1]	1 [1, 1]	0 [0, 0]	0 [0, 0]	0 [0, 0]	0.4993	Not worth bet on
Fairness 1	1 [0.9, 1]	1 [1, 1]	0 [−0.1, 0]	0.1 [0, 0.2]	0 [0, 0]	0.0718	Promising but risky bet against
Fairness 2	1 [1, 1]	1 [1, 1]	0 [0, 0]	0 [0, 0]	0 [0, 0]	0.5016	Not worth bet on

As explained in Section 3.5, *QI* 1 to 5 correspond to each question regarding integrity and *QI Mean* is their average. Honesty, Commitment, Fairness 1 and 2 are the observed metrics related to integrity.

We have also included Figures that illustrate the values in the tables. In particular, we have picked one figure that shows the correlation between one subjective and one observed metrics, such as the case of QB Mean (benevolence questions) and Favouritism in Figure 3. We also show how the correlation between **subjective trustworthiness** (**STW**) and observed trustworthiness, *TW abs* and *TW rel*, respectively, in Figures 4 and 5. In these figures, we can see the actual data in the red bar charts on the axis and black circles (each representing an instance) in the middle of the plot. The blue lines around the red charts show the credible normal distributions for our data. As explained in the beginning of this section, the correlation is then run for the several credible

Table 5. Bayesian Pearson Correlation Between Subjective and Observed Metrics, Presenting the Means of the Correlation Among the Distributions of Two Metrics, the Probability of this Mean Being Positive and the Evaluation According to Reference [18]

Metric 1	Metric 2	Mean	%	Evaluation
STW	TW abs	0.3 [0, 0.5]	0.9855	Good bet - too good to disregard
STW	TW rel	0.5 [0.2, 0.7]	0.9999	Nearing certainty
QA Mean	Time/Task	−0.2 [−0.5, 0.1]	0.1035	Casual bet against
QA Mean	Moves/Task	0 [−0.3, 0.3]	0.5117	Not worth bet on
QA Mean	Moves/Time	0.2 [−0.1, 0.4]	0.8590	Casual bet
QB Mean	Favouritism	0.3 [0, 0.5]	0.9802	Good bet - too good to disregard
QI Mean	Honesty	0.1 [−0.2, 0.4]	0.7757	Casual bet
QI Mean	Commitment	0 [−0.3, 0.3]	0.4976	Not worth bet against
QI Mean	Fairness 1	0.2 [−0.1, 0.5]	0.8692	Casual bet
QI Mean	Fairness 2	0 [−0.3, 0.3]	0.5087	Not worth bet on

In particular, we present the correlations between (1) STW, which is an average of all the questions in the questionnaire (i.e., QA, QB and QI), and observed trustworthiness, both absolute and relative, i.e., TW abs and TW rel, respectively; (2) subjective metric of ability (QA Mean, which is the average of ability questions) and objective metrics of ability (Time/Task,Moves/Task and Moves/Time); and in a similar fashion for metrics of (3) benevolence and (4) integrity.

Table 6. Bayesian Pearson Correlation Between Metrics of Overall Trustworthiness and the Observed Metrics of Each Aspect of Ability, Benevolence, and Integrity

Metric 1	Metric 2	Mean	%	Evaluation
TW rel	Time/Task	−0.1 [−0.4, 0.2]	0.2844	Not worth bet against
TW rel	Moves/Task	0.3 [0, 0.5]	0.9740	Good bet - too good to disregard
TW rel	Moves/Time	0.2 [0, 0.5]	0.9451	Promising but risky bet
TW rel	Favouritism	0.8 [0.6, 0.9]	0.9999	Nearing certainty
TW rel	Honesty	0.3 [−0.1, 0.5]	0.9416	Promising but risky bet
TW rel	Commitment	0 [−0.3, 0.3]	0.4979	Not worth bet against
TW rel	Fairness 1	0.3 [0, 0.5]	0.9583	Good bet - too good to disregard
TW rel	Fairness 2	0 [−0.3, 0.3]	0.4975	Not worth bet against
TW abs	Time/Task	−0.7 [−0.8, −0.5]	0.0001	Promising but risky bet against
TW abs	Moves/Task	0 [−0.3, 0.3]	0.4630	Not worth bet against
TW abs	Moves/Time	0.7 [0.6, 0.8]	0.9999	Nearing certainty
TW abs	Favouritism	0.2 [0, 0.5]	0.9411	Promising but risky bet
TW abs	Honesty	0.2 [−0.1, 0.5]	0.8778	Casual bet
TW abs	Commitment	0 [−0.3, 0.3]	0.5009	Not worth bet on
TW abs	Fairness 1	0.1 [−0.2, 0.4]	0.7863	Casual bet
TW abs	Fairness 2	0 [−0.3, 0.3]	0.5047	Not worth bet on

The table presents the means of the correlation among the distributions of two metrics, the probability of this mean being positive and the evaluation according to [18].

distributions, creating a distribution of possible correlation values. On top of the figures, we can see the distribution of the correlation possible values and their 95% HDI. The more similar the credible distributions of the data are, the slimmest the correlation distribution will be and the more likely it is that the correlation is the median value. The furthest this distribution is from 0, the more probable it is that there is indeed a correlation, positive (if on the positive side) or negative (if on the negative side).

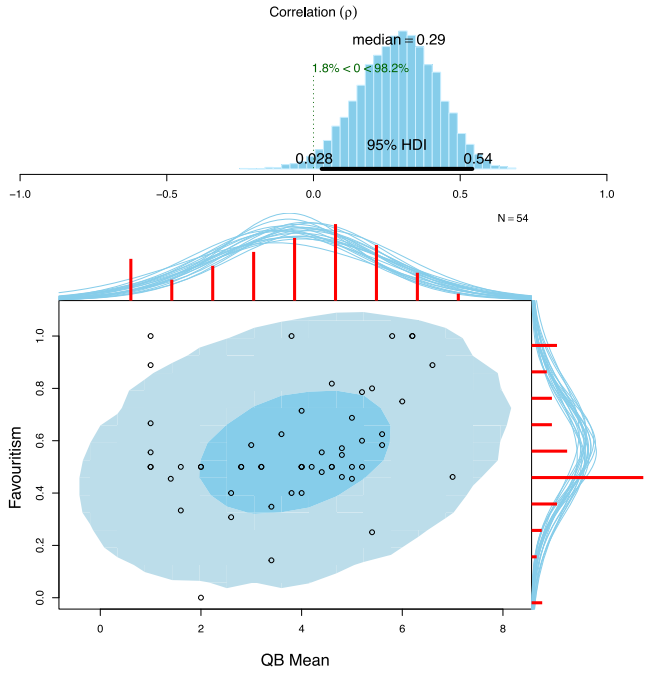

Fig. 3. Bayesian Pearson Correlation between subjective benevolence (QB Mean) and observed benevolence (Favouritism). The black circles represent the instances of our dataset, and the metrics' histograms are on the red bar charts. The blue lines around the red charts show the credible normal distributions for our data. The top plot presents the distribution of the correlation of possible values and their 95% HDI.

4.4 Strategy

In Table 7, we can see the results of the multiple choice question regarding participant's strategy when performing the task. We have ordered the table according to the original order in the questionnaire and numbered them for reference simplicity. In the complementing text of "Other", 3 participants reported "slightly" helping more agent X, whereas 1 reported similar regarding agent Z.

5 RESULTS DISCUSSION

In this section, we discuss the results, highlight some interesting aspects and reflect on what they might mean. We divide this section similarly to the result section to make it easier to follow. However, we find it relevant to highlight our main findings beforehand, as some of these findings affect the interpretation of the several results presented. As such, our main findings looking at the numbers were:

(1) There were expected differences between conditions both for self-reported (subjective) and observed metrics, meaning our manipulations had effect on the participants.
(2) STW highly correlates with both observed overall consequence of trustworthiness metrics (TW abs and TW rel).
(3) There are probable yet low correlations (probable here means that after several correlation tests with the possible distributions of the data, the correlation was mostly either positive or negative, making it a safer bet), between subjective and one of the observed metrics for each of the aspects of ability, benevolence and integrity.

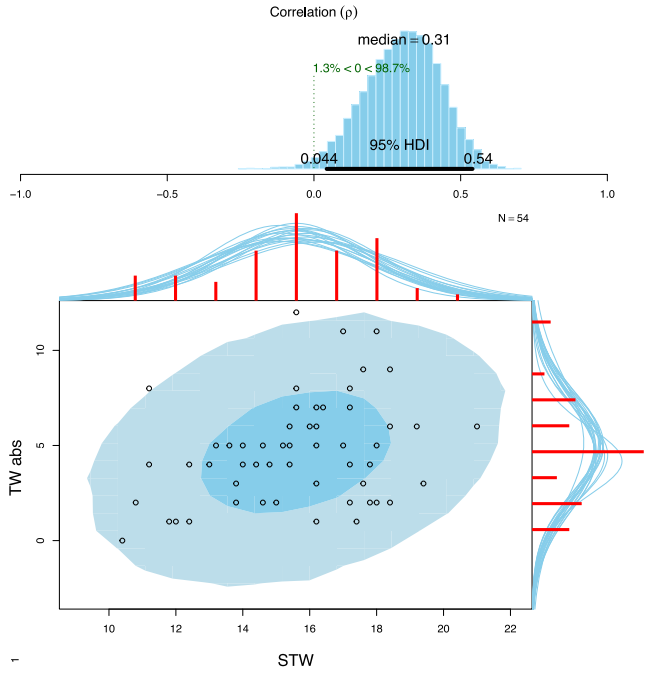

Fig. 4. Bayesian Pearson Correlation between STW and observed absolute trustworthiness (TW abs). The black circles represent the instances of our dataset, and the metrics' histograms are on the red bar charts. The blue lines around the red charts show the credible normal distributions for our data. The top plot presents the distribution of the correlation of possible values and their 95% high density interval.

Table 7. Number of Times Each *goal* was Chosen in Strategy Question (Multiple Choice Allowed), Ordered as in Questionnaire

Option	#
1. Collect as many products as possible	39
2. Collect products as fast as possible	34
3. Maximize team score	26
4. Maximize personal score	16
5. Collect easiest products (based on icon)	26
6. Collect easiest products (based on distance)	25
7. Collect according to the chat messages	3
8. Helping both agents equally	15
9. Helping specifically agent X	2
10. Helping specifically agent Z	2
11. I do not know what my goal was	1
12. Other	3

(4) Almost all observed metrics of ability, benevolence and integrity correlated with TW rel and most of them also correlated with TW abs. These correlations were rather low except for Favouritism (benevolence) with TW rel and Time/Task and Moves/Time with TW abs, which were expected by definition.

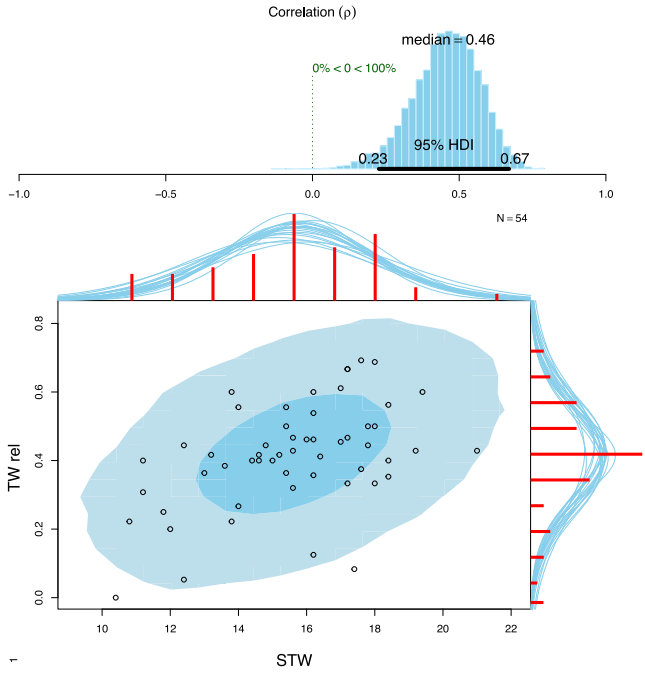

Fig. 5. Bayesian Pearson Correlation between STW and observed relative trustworthiness (TW rel). The black circles represent the instances of our dataset, and the metrics' histograms are on the red bar charts. The blue lines around the red charts show the credible normal distributions for our data. The top plot presents the distribution of the correlation of possible values and their 95% HDI.

(5) Most participants said they collected as many products as possible and as fast as possible. They also mentioned they chose the easiest products to collect. This can be interpreted as participants caring mostly about doing the task well and quickly, with the least effort associated to it, and not paying so much attention to the details of the scenario.

Beyond the numbers, we found several interesting points transversal to all results. These were:

(1) Most participants only cared about the game/task itself, focusing on performing well at finding and retrieving objects rather than paying attention to the context that was presented initially (i.e., participants tend to forget that one of the agents was their friend, in benevolence condition, or that it is their last day of work, and they need to make as much money as possible, in integrity condition). They also rarely looked at the chat, from our observation.
(2) Probably related to the previous point, some participants mentioned that they do not see the agents as their teammates. This may mean that our task could be improved by including, for example, more interdependencies or autonomy.
(3) Very few participants lied or gave up on tasks. This may be because they feel observed (which affects ecological validity) or because they do not feel compelled to lie/give up in such scenario.
(4) It can be quite challenging to self-report measures of one's ability, benevolence and integrity, specially without a term of comparison. As such, we need to take these metrics' results critically.
(5) Some participants verbalized that they think it does not make sense for us to ask them the questions related to benevolence (we can see in Figure 3 that there were several low average

scores). This reflects that many of them did not feel like they were helping or "having the back" of Agent X, for example. However, results tend to show that some participants did feel that way, as there were also higher scores, different between groups, and a correlation with Favouritism.

In the remaining of this section, we will go over each table and analyse the results in more detail, keeping in mind the points stated above. At the end, we will compare our main findings with the purpose of this study.

5.1 Differences Between Groups

It is interesting to see that all questions regarding self-reported benevolence (Table 2 show higher scores in the benevolence group (Gb) than in the normal group (Gn), all of these with high probabilities associated. In particular, QB3, QB4, and QB5 present the highest probabilities of in fact differing between the two groups showing a "promising but risky bet" according to Reference [18], with mean differences higher than 1 (in a 7-point scale). These questions are about having *"worked to protect Agent X"*, *"watched teammate Agent X's back (synonym: to look out for Agent X in case it needs assistance)"* and *"looked out for teammate Agent X"*. It is worth mentioning that the question that presents highest probability of difference (QB4) is the one that translates better to our chosen metric for observed benevolence, i.e., give more assistance to Agent X than Agent Z. However, Favouritism presents only a casual bet that it differs between Gb and Gn. This may mean that although our narrative made participants care more about Agent X (as a feeling), there was not a relevant difference in how much they actually helped one agent over the other (in action).

When comparing Ga with Gn (Table 3, we see that it is promising to bet that there is a difference in answer to QA1, which is "As a teammate, I was capable at my jobs", with a difference of 0.7 between the groups' means. The other questions, however, do not show big differences among the two groups, which leads to only a casual bet in the QA Mean comparison. In general it is hard to have participants evaluate themselves regarding ability, since they do not know how other participants performed. In particular, how good they actually were in the task might have been evenly distributed among groups. We cannot know for sure if this was the case, but we can see that Moves/Time, which indicates how fast participants were in general, should in that case not differ between groups and it did not. If they were evenly distributed in terms of actual ability, that might have made the manipulation of the group imperceptible for the participants in terms of how much easier the task becomes. We did expect, however, that it would show a difference in the observed metrics, such as Time/Task and Moves/Task, since the participants in Ga group could see at all times when the product they should collect was, and in fact both suggest a casual bet (it is against as the higher the time taken to perform a task, the lower expected ability). Interestingly, we can see that the standard deviations in Ga are much lower than in Gn for Moves/Task and Time/Task, which probably means that our manipulation created less difference among participants of Ga regarding these two metrics. This makes sense as all participants in the group could see all the products from afar, making the task simply easier for everyone, making their actual differences less significant.

Finally, although Gi participants seem to score lower in self-reported subjective measures (QI metrics) than in Gn, these are not very relevant in terms of Mean differences and percentage. This is understandable when we look at the observed metrics, which mostly show either 0 or 1, indicating that there were very few people lying and giving up. This means that the manipulation in terms of integrity was not successful.

Although we can see there were expected differences between the groups, showing that we did manipulate the participants in a certain way, results show that our manipulations did not work for

all subjective (self-reported) and observed metrics. We speculate this happens because of several reasons observed during the experiment (already stated in the beginning of this section).

5.2 Correlation Between Subjective and Observed Metrics

In Table 5, we can find the Bayesian Pearson correlations between subjective and objective metrics. Self-reported STW highly correlates with observed relative trustworthiness (TW rel) near certainty, and it also a good bet that it positively correlates with absolute trustworthiness (TW abs). This means that overall participants answered the questionnaire according to their performance in the task and help towards Agent X. These results align with the very likely (98%) positive correlation between both self-reported subjective and observed metrics for benevolence. The subjective metrics of ability also seem to possibly (only a casual bet) correlate with two of the observed metrics of ability, such as Moves/Time and Time/Task, although these would not be very high correlations. Moves/Task, however, seems to not correlate at all. It may be that because of the nature of the task, Moves/Task is not a good measure of ability. In our task, participants usually choose to go through all the corridors until they find the product they need to collect. We expected some participants to remember the corridors better than others and that would differ them in terms of average Moves/Task. Although we cannot tell based on this correlation only that the measure did not succeed in capturing participants' ability, it probably did not suit our task. Finally, as said before, very few participants lied, it seems that their self-reported metrics of integrity possibly (87%) correlate with observed metric Fairness 1 (low correlation). Measures of commitment and fairness w.r.t. commitment (Fairness 2) do not correlate at all with self-reported Integrity. As we saw in the previous section almost no participant gave up in the task, which makes it impossible to evaluate these metrics.

Even though it is possible that most subjective metrics correlate with objective metrics in the expected direction, most of these correlations are not high. The reasons for this may be (1) because the self-reported measures do not actually represent the participants' trustworthiness aspects of ability, benevolence and integrity or (2) because the observed metrics are not sufficient to capture the participants' trustworthiness aspects of ability, benevolence, and integrity. Ideally, we would compare these with a ground-truth values of ability, benevolence, and integrity, but unfortunately there is no way we can have these. We can, however, see how the observed metrics of each aspect relate to overall metrics of trustworthiness.

5.3 Correlation Between Observed Trustworthiness and Other Observed Metrics

The observed metrics for trustworthiness are by definition related with the observed metrics for each of the aspects of ability, benevolence and integrity. For example, the number of successes (which is taken into account for both absolute and relative trustworthiness, i.e., TW abs and TW rel, respectively) is related to how fast a participant can do a task (Time/Task) and also how many times the participant decided to help Agent X instead of Agent Z (Favouritism), and so on. In this subsection, we analyse how actually these metrics correlate (in Table 6). We can see that TW rel (which can be interpreted as how likely it is for the participant to collaborate with Agent X) probably positively correlates with Moves/Time (95%) and Moves/Task (97%), Favouritism ($\tilde{1}$00%), Honesty (94%) and Fairness 1 (96%). In particular, the correlation between TW rel and Favouritism is extremely high (0.8), which is expected given their definitions, i.e., percentage of presented tasks by Agent X that were successful and percentage of all successful tasks that were for Agent X. As for TW abs (which can be interpreted with how much a participant actually helped Agent X in absolute terms), we can see that it highly certainly ($\tilde{1}$00%) highly correlates with Moves/Time and negatively with Time/Task ($\tilde{1}$00%, as it shows $\tilde{0}$% for positive correlation). There is a high chance

that it slightly correlates with Favouritism (94%) and, not as probable, with Honesty (88%). Again, commitment and Fairness 2 do not correlate at all, probably given to the scarcity of these metrics.

5.4 Strategy

In Table 7, we have reported the results for the question related to strategy. We realized that many times, strategy was based on the least effort for the participant, either by having already seen that product icon before or because it was simply closer to them. Supported by literature [10, 65] and our results, we speculate we should consider human's *cost-benefit evaluation*, i.e., participants choose whether or not the reward is worth the perceived effort, and this affects their decision. In particular, the law of least effort plays a central role in decision-making, i.e., when presented to two tasks with equal rewards, one will choose the least effortful [58]. This may mean that more than paying attention to the task context in terms of benevolence and integrity, participants might have been in fact playing according to their own perceived benefit with the lowest effort possible. When determining how much someone should be trusted, then, it may be more helpful to know what is for them a risk/effort and reward/benefit. This may also be helpful to predict where the human teammates may do next, once we detect the strategy they are using (for example, going for the icons they have already seen before).

5.5 Implications for Human Trustworthiness in Human-AI Teams

In this article, we wanted to study what makes a human trustworthy in a human-AI team setting, and how could an artificial agent observe it, given a specific task. Although we knew we would not be able to prove it (due to lack of ground-truth), we wanted to explore how cues on ability, benevolence and integrity could be used as observable trustworthiness. The goal was to take the first step towards understanding how an artificial agent can form a belief of trustworthiness regarding its human teammate. In particular, we wanted to explore how an agent could form beliefs on (1) whether the human could do a task and on (2) whether they would do it. In light of these results and the main purpose of this study, we believe that participants' ability, benevolence and integrity do affect their overall trustworthiness and the direct consequences of it, but may not be the best human's krypta in human-AI teamwork scenario, given low correlations and overall findings, as we discuss in this subsection.

In terms of manifesta, the cue of Favouritism highly correlated with whether the human did the task the agent asked for, and cues of performance (w.r.t. moves) and commitment seem to have also affected this. Similarly, cues of performance w.r.t. to both time and moves highly correlate with how well they would do it. However, the cues of integrity did not seem to be suitable for this environment. Although these seem like promising krypta and manifesta, we suspect that this may not be the best model for this type of task and environment. In particular, both benevolence and integrity definitions and cues may not be appropriate, as most participants did not feel particularly inclined towards teaming up with (and helping) one agent or giving up/lying.

From our understanding, there might be better ways of detecting the willingness to perform a task successfully in human-AI team settings, particularly in short interactions, such as detecting overall strategy or personal preferences. As we suspected during the pilot study, participants were mostly paying attention to the task itself and to how to solve it efficiently, instead of caring about the interactions with the agents or even getting points by just lying. Although this may have to do mostly with our setup, we believe that in order to understand whether a person will do something, we need to understand how, in general, the person is executing tasks (what is their strategy). This may depend on, for example, what the execution of the task represents in terms of risk and reward for that person. And, of course, risk and reward may be related to a person's ability, benevolence and integrity. It may also be related with something else, though, such as a personal preference.

After this experiment, we also consider that different tasks may require different manifesta and krypta apply. Overall, these measures must be further explored in different scenarios, as well as compared with other possible trustworthiness models as krypta (and respective manifesta).

5.6 Limitations

The setting in which we ran the experiment presented some technical difficulties. The server presented a lag for all participants, the higher the further from it they were, which might have negatively affected our results. Moreover, the setting of the experiment might have made participants feel too "observed", having slightly harmed the ecological validity of the study (possibly affecting the give-ups and lies, as discussed in the beginning of this section). The task in our experiment was a short one (10min) and did not involve high interdependence, which may have contributed to the feeling of not being teammates. These characteristics of the task may also not be the most suitable to benevolence and integrity. For such tasks, it may be more relevant to look at other models of trust, such as swift trust, see e.g., References [25, 31]. In fact, future research should also consider that if trust changes overtime, so can trustworthiness.

Overall, It is still an open question how to appropriately model the willingness of the human teammate to perform an action, as such question is also not yet answered by social sciences. In particular, it would be interesting to explore measures that do not assume complete knowledge (i.e., measures that take into account only interactions with one agent). What's more, our task did not allow the participants to simply fail due to lack of ability. This is justified in Section 3.2.1, but it might be interesting to explore a task where it is possible to (1) lie, (2) give up and (3) fail (due to lack of ability), while being possible to tell these 3 apart.

Another limitation of our work mentioned before is the fact that there is no ground-truth or baseline we can compare our results with. As such, we have proposed observed metrics for the aspects we relate with trustworthiness (ability, benevolence and integrity) and two observed consequences of trustworthiness metrics (TW rel and TW abs), but we cannot test them. We can only compare them with subjective metrics which we are not certain about how well they capture what we want to know (i.e., how trustworthy someone is), given its dependency on self and context awareness of the participants. We look forward to exploring human trustworthiness in different scenarios and with different tasks and more sophisticated agents, in order to compare with and improve our current model.

5.7 Future Directions

Our future directions include exploring the manifesta and krypta further, their developments and context dependencies in time and throughout interactions. We will work on using artificial trust beliefs for decision-making support, both for autonomous decision-making of the artificial teammate and for human support. It is also important to explore ways of evaluating artificial trust models, e.g., by creating test-beds for such experiments. Further explanation of the next steps of our work can be found in Reference [15].

In this work, we see that both benevolence and integrity probably affect human actions in human-AI teamwork and should be taken into account when designing such teams. However, benevolence and integrity may not be the most direct aspects to either measure or use for prediction. On the other hand, as we observed, there likely is an influence of participants' strategy on which task they choose and whether they succeed on it. As such, we want to work on modelling willingness (i.e., factors other than competence that contribute to the success of the task) more concretely, e.g., using aspects which are more task and context dependent, such as possibility, interest in doing a certain type of task, preference, disposition, and so on). We want to use these aspects to build a model for informed and justified decision-making of the artificial teammate, making use of

principles of Interdependence Analysis and Coactive Design [28]. Such tool can also help human decision-making in human-AI teamwork scenarios, by providing an overview of the feasibility of different team configurations, for example. Furthermore, we want to update this decision-making model through interaction using existing models such as Reference [5]. In future research, we want to use scenarios where collaboration is more necessary and explicit. Ideally, these scenarios also provide different levels of interdependence which allow the teammates to choose whether to engage in certain tasks or whether to help a teammate, for example, and can include mixed-motives. Such scenarios include search and rescue tasks, as in Reference [62], moving-out tasks, as in Reference [14], or even cooking tasks, as in Reference [36].

Human's trust and trustworthiness are also affected by the artificial teammate's behaviour, see e.g., Reference [14]. It would also be interesting to investigate the human's trustworthiness in situations where the artificial teammate behaves differently. For example, situations where the artificial teammate does not obey to its immediate human teammate have been recently studied, see e.g., References [53, 59]. Such situations can have mixed motives and knowledge, e.g., the human may want to go straight whereas the artificial teammate knows that it is dangerous (e.g., the human does not have the skills required for what's ahead), so it opposes. We would like to explore how human's trust and trustworthiness change in such situations and how they depend on the outcome, i.e., if it turns out that the agent is right or wrong has any impact on the collaboration and human trustworthiness (for example, changing the human's willingness).

6 CONCLUSION

In this article, we take a first step towards implementing beliefs of artificial trust through interaction. According to the previously presented model of human trustworthiness in human-AI teams, we proposed a set of metrics for observable human behaviours (manifesta), representing aspects of their trustworthiness through their krypta (in this case ability, benevolence and integrity). Both theoretically and empirically, we have explored these metrics and compared them with self-reported subjective metrics of the same krypta. We have also compared both observed and self-reported metrics with overall metrics of trustworthiness (based on absolute and relative task success, from the perspective of one agent). Results showed that there was a high correlation between the average of the self-reported subjective metrics and observed metrics of overall trustworthiness. However, when dividing these metrics into ability, benevolence, and integrity, the subjective-observed pair-wise correlations were not so high, which may mean that although these metrics are relevant they may not correspond well to each of the aspects. We can see that ability, benevolence and integrity do affect their overall trustworthiness and the direct consequences of it, but may not be the best human's krypta in human-AI teamwork scenario. On the other hand, we observed that, the human teammate's decision on whether to perform a task depended on a strategy, related to participant's goals and cost-benefit analysis. Unfortunately, we cannot have ground-truth or other models to compare our results with, but these results shed light on how different aspects of trustworthiness can be used for human behaviour prediction in human-AI teamwork scenario. Although there is a need for further exploration of these metrics, this article presents an important step towards building an intelligent agent capable of building beliefs of trust in human teammates and therefore capable of making informed decisions to achieve the team's goal.

CODE AND DATA AVAILABILITY

The code used for the task presented in this article can be found in https://github.com/centeio/click-collect. The resulting data, as well as the questionnaire used, can be found in https://doi.org/10.4121/21982991.v1

ACKNOWLEDGMENTS

Thanks to all colleagues who helped us find the right methods and analysis for this experiment, and of course to those who proofread the manuscript.

REFERENCES

[1] Barbara D. Adams, Sonya Waldherr, and J. Sartori. 2008. Trust in teams scale, trust in leaders scale: Manual for administration and analyses. https://cradpdf.drdc-rddc.gc.ca/PDFS/unc95/p530364_A1b.pdf

[2] Barbara D. Adams and R. Webb. 2002. Trust in small military teams. *Command and Control Research Program* (2002). https://apps.dtic.mil/sti/pdfs/ADA630700.pdf

[3] Nele Albers, Mark A. Neerincx, and Willem-Paul Brinkman. 2022. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active. *Plos One* 17, 12 (2022). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0277295

[4] Nele Albers, Mark A. Neerincx, and Willem-Paul Brinkman. 2022. Addressing people's current and future states in a reinforcement learning algorithm for persuading to quit smoking and to be physically active: Data and analysis code. Version 2. 4TU.ResearchData. dataset.

[5] Arsha Ali, Hebert Azevedo-Sa, Dawn M. Tilbury, and Lionel P. Robert Jr. 2022. Heterogeneous human–robot task allocation based on artificial trust. *Scientific Reports* 12, 1 (2022), 15304.

[6] Hebert Azevedo-Sa, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. 2021. A unified bi-directional model for natural and artificial trust in human-robot collaboration. *IEEE Robotics and Automation Letters* 6, 3 (2021), 5913–5920. DOI: https://doi.org/10.1109/LRA.2021.3088082

[7] Rasmus Bååth. 2014. Bayesian first aid: A package that implements bayesian alternatives to the classical *.test functions in R. In *Proceedings of the UseR! 2014—the International R User Conference.*

[8] Michael Bacharach and Diego Gambetta. 2001. *Trust as Type Detection.* Springer Netherlands, Dordrecht, 1–26. DOI: https://doi.org/10.1007/978-94-017-3614-5_1

[9] D. v.d. Bergh, M. A. Clyde, A. R. K. N. Gupta, et al. 2021. A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behav Res* 53 (2021), 2351–2371.

[10] M. M. Botvinick and Z. B. Rosen. 2009. Anticipation of cognitive demand during decision-making. *Psychol Res.* 73, 6 (2009), 835–42. DOI: 10.1007/s00426-008-0197-8

[11] Matthias Brand, Kirsten Labudda, and Hans J. Markowitsch. 2006. Neuropsychological correlates of decision-making in ambiguous and risky situations. *Neural Networks* 19, 8 (2006), 1266–1276.

[12] Christina Breuer, Joachim Hüffmeier, Frederike Hibben, and G. Hertel. 2020. Trust in teams: A taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams. *Human Relations* 73 (2020), 3–34.

[13] Lucile Callebert, Domitile Lourdeaux, and Jean-Paul A. Barthès. 2016. A trust-based decision-making approach applied to agents in collaborative environments. In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016), Volume 1, Rome, Italy, February 24-26, 2016.* H. Jaap van den Herik and Joaquim Filipe (Eds.), SciTePress, 287–295. DOI: https://doi.org/10.5220/0005825902870295

[14] Carolina Centeio Jorge, Nikki Helena Bouman, Catholijn Jonker, and Myrthe Lotte Tielman. 2023. Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. *Frontiers in Robotics and AI* 10 (2023), 1143723. https://www.frontiersin.org/articles/10.3389/frobt.2023.1143723

[15] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2023. Artificial trust for decision-making in human-AI teamwork: Steps and challenges. In *Proceedings of the HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI).*

[16] Carolina Centeio Jorge, Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. Trust should correspond to trustworthiness: A formalization of appropriate mutual trust in human-agent teams. In *Proceedings of the International Workshop in Agent Societies.*

[17] Carolina Centeio Jorge, Myrthe L. Tielman, and Catholijn M. Jonker. 2022. Assessing artificial trust in human-agent teams: a conceptual model. In *IVA'22: ACM International Conference on Intelligent Virtual Agents, Faro, Portugal, September 6–9, 2022,* Carlos Martinho, João Dias, Joana Campos, and Dirk Heylen (Eds.), ACM, 24:1–24:3. DOI: https://doi.org/10.1145/3514197.3549696

[18] R.A. Chechile. 2020. *Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution-Free Methods.* MIT Press.

[19] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (1951), 297–334.

[20] Mohammadreza Esfandiari, Senjuti Basu Roy, and Sihem Amer-Yahia. 2018. Explicit preference elicitation for task completion time. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* 1233–1242.

[21] Rino Falcone and Cristiano Castelfranchi. 2004. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proceedings of the AAMAS*. IEEE Computer Society, 740–747. DOI: https://doi.org/10.1109/AAMAS.2004.10084

[22] Rino Falcone, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. 2013. From manifesta to krypta: The relevance of categories for trusting others. *ACM Transactions on Intelligent Systems and Technology* 4, 2 (3 2013). DOI: https://doi.org/10.1145/2438653.2438662

[23] Qianrao Fu, Herbert Hoijtink, and Mirjam Moerbeek. 2021. Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods* 53, 1 (2021), 139–152.

[24] David Gefen, Elena Karahanna, and Detmar W. Straub. 2003. Trust and TAM in online shopping: An integrated model. *MIS Q.* 27, 1 (2003), 51–90.

[25] Kerstin S. Haring, Elizabeth Phillips, Elizabeth H. Lazzara, Daniel Ullman, Anthony L. Baker, and Joseph R. Keebler. 2021. Applying the swift trust model to human-robot teaming. In *Proceedings of the Trust in Human–Robot Interaction*. Elsevier, 407–427.

[26] Michael Inzlicht, Amitai Shenhav, and Christopher Y. Olivola. 2018. The effort paradox: Effort is both costly and valued. *Trends in Cognitive Sciences* 22, 4 (2018), 337–349. DOI: https://doi.org/10.1016/j.tics.2018.01.007

[27] Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman, and Md Abdullah Al Fahim. 2018. Initial trustworthiness perceptions of a drone system based on performance and process information. In *Proceedings of the 6th International Conference on Human-Agent Interaction, HAI 2018, Southampton, United Kingdom, December 15-18, 2018*. Michita Imai, Tim Norman, Elizabeth Sklar, and Takanori Komatsu (Eds.). ACM, 229–237. DOI: https://doi.org/10.1145/3284432.3284435

[28] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. Van Riemsdijk, and M. Sierhuis. 2014. Coactive design: designing support for interdependence in joint activity. *Journal of Human-Robot Interaction* 3, 1 (2014), 43–69.

[29] Matthew Johnson and Jeffrey M. Bradshaw. 2021. The role of interdependence in trust. In *Proceedings of the Trust in Human–Robot Interaction*. Elsevier, 379–403.

[30] Matthew Johnson, Catholijn M. Jonker, M. Birna van Riemsdijk, Paul J. Feltovich, and Jeffrey M. Bradshaw. 2009. Joint activity testbed: Blocks world for teams (BW4T). In *Proceedings of the Engineering Societies in the Agents World X*, Vol. 5881. Springer, 254–256. DOI: https://doi.org/10.1007/978-3-642-10203-5_26

[31] Lionel P. Robert Jr., Alan R. Dennis, and Yu-Ting Caisy Hung. 2009. Individual swift trust and knowledge-based trust in face-to-face and virtual team members. *J. Manag. Inf. Syst.* 26, 2 (2009), 241–279. DOI: https://doi.org/10.2753/mis0742-1222260210

[32] Monika Kaczorowska, Paweł Karczmarek, Małgorzata Plechawska-Wójcik, and Mikhail Tokovarov. 2021. On the improvement of eye tracking-based cognitive workload estimation using aggregation functions. *Sensors* 21, 13 (2021), 4542.

[33] Esther S. Kox, José H Kerstholt, Tom F. Hueting, and Peter W. de Vries. 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems* 35, 2 (2021), 30.

[34] John K. Kruschke. 2013. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General* 142, 2 (2013), 573.

[35] Daniella Laureiro-Martínez, Stefano Brusoni, and Maurizio Zollo. 2010. The neuroscientific foundations of the exploration- exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics* 3, 2 (2010), 95.

[36] Marin Le Guillou, Laurent Prévot, and Bruno Berberian. 2023. Trusting artificial agents: Communication trumps performance. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 299–306.

[37] John D. Lee and Katrina A. See. 2004. Trust in automation: designing for appropriate reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society* 46 (2004), 50–80.

[38] Michael Lewis, Huao Li, and Katia Sycara. 2020. Deep learning, transparency, and trust in human robot teamwork. In *Proceedings of the Trust in Human–Robot Interaction*. Elsevier, 321–352.

[39] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. *The Role of Trust in Human–Robot Interaction*. Springer International Publishing. 135–159 pages. DOI: https://doi.org/10.1007/978-3-319-64816-3_8

[40] Michael Luck and Mark D'Inverno. 1996. Engagement and cooperation in motivated agent modelling. In *Distributed Artificial Intelligence Architecture and Modelling*. Chengqi Zhang and Dickson Lukose (Eds.). Springer, Berlin, 70–84.

[41] Roger C. Mayer and James H. Davis. 1999. The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology* 84, 1 (1999), 123.

[42] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Source: The Academy of Management Review* 20, 3 (1995), 709–734.

[43] Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

[44] D. Harrison McKnight, Vivek Choudhury, and Charles J. Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Inf. Syst. Res.* 13, 3 (2002), 334–359. DOI: https://doi.org/10.1287/isre.13.3.334.81

[45] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More similar values, more trust? - the effect of value similarity on trust in human-agent interaction. In *AIES'21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*. Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, 777–783. DOI: https://doi.org/10.1145/3461702.3462576

[46] A. M. Mohamed and M. N. Huhns. 2001. Multiagent benevolence as a societal norm. In *Social Order in Multiagent Systems. Multiagent Systems, Artificial Societies, and Simulated Organizations*, R. Conte and C. Dellarocas (Eds.). Vol 2. Springer, Boston, MA.

[47] Changjoo Nam, Phillip Walker, Huao Li, Michael Lewis, and Katia Sycara. 2020. Models of trust in human control of swarms with varied levels of autonomy. *IEEE Transactions on Human–Machine Systems* 50, 3 (6 2020), 194–204. DOI: https://doi.org/10.1109/THMS.2019.2896845

[48] C. Nass and Y. Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56, 1 (2000), 81–103.

[49] Shofiyati Nur Karimah, Teruhiko Unoki, and Shinobu Hasegawa. 2021. Implementation of long short-term memory (LSTM) models for engagement estimation in online learning. In *Proceedings of the 2021 IEEE International Conference on Engineering, Technology Education (TALE)*. 283–289. DOI: https://doi.org/10.1109/TALE52509.2021.9678909

[50] James Onken, Reid Hastie, and William Revelle. 1985. Individual differences in the use of simplification strategies in a complex decision-making task. *Journal of Experimental Psychology: Human Perception and Performance* 11, 1 (1985), 14.

[51] Michael E. Palanski and Francis J. Yammarino. 2007. Integrity and leadership: Clearing the conceptual confusion. *European Management Journal* 25, 3 (2007), 171–184.

[52] Daniel S. Quintana and Donald R. Williams. 2018. Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP. *BMC Psychiatry* 18, 1 (2018), 1–8.

[53] R. Rachum, Y. Nakar, and R. Mirsky. 2023. Stubborn: An environment for evaluating stubbornness between agents with aligned incentives. *arXiv preprint arXiv:2304.12280*. https://doi.org/10.48550/arXiv.2304.12280

[54] Ellen Rusman, Jan Van Bruggen, Peter Sloep, and Rob Koper. 2010. Fostering trust in virtual project teams: Towards a design framework grounded in a TrustWorthiness ANtecedents (TWAN) schema. *International Journal of Human-Computer Studies* 68, 11 (2010), 834–850.

[55] Jordi Sabater-Mir and Laurent Vercouter. 2013. Trust and reputation in multiagent systems. *Multiagent Systems* (2nd edition). 381–401.

[56] Eduardo Salas, Dana, E. Sims, and C. Shawn Burke. 2005. Is there a "Big Five" in teamwork? *Small Group Research*, 36, 5 (2005), 555–599. DOI: 10.1177/1046496405277134

[57] Johannes Schiebener and Matthias Brand. 2015. Self-reported strategies in decisions under risk: Role of feedback, reasoning abilities, executive functions, short-term-memory, and working memory. *Cognitive Processing* 16, 4 (2015), 401–416.

[58] R. L. Solomon. 1948. The influence of work on behavior. *Psychological Bulletin* 45, 1 (1948), 1–40.

[59] Kavyaa Somasundaram, Andrey Kiselev, and Amy Loutfi. 2023. Intelligent disobedience: A novel approach for preventing human induced interaction failures in robot teleoperation. In *Companion of the 2023 ACM/IEEE International Conference on Human–Robot Interaction, HRI 2023, Stockholm, Sweden, March 13-16, 2023*. Ginevra Castellano, Laurel D. Riek, Maya Cakmak, and Iolanda Leite (Eds.). ACM, 142–145. DOI: https://doi.org/10.1145/3568294.3580060

[60] Vidullan Surendran, Kasra Mokhtari, and Alan R. Wagner. 2021. Your robot is watching 2: Using emotion features to predict the intent to deceive. In *30th IEEE International Conference on Robot & Human Interactive Communication, RO-MAN 2021, Vancouver, BC, Canada, August 8-12, 2021*, IEEE, 447–453. DOI: https://doi.org/10.1109/RO-MAN50785.2021.9515553

[61] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. 2013. A socio-cognitive perspective of trust. In *Proceedings of the Agreement Technologies*. Springer, 419–429.

[62] Ruben S Verhagen, Mark A. Neerincx, Can Parlar, Marin Vogel, and Myrthe L. Tielman. 2023. Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2316–2318.

[63] Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. 2019. Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B: Biological Sciences* 374, 1771 (4 2019). DOI: https://doi.org/10.1098/rstb.2018.0032

[64] E. J. de Visser, M. M. M. Peeters, M. F. Jung, et al. 2020. Towards a theory of longitudinal trust calibration in human -robot teams. *Int J. of Soc Robotics* 12 (2020), 459–478. https://doi.org/10.1007/s12369-019-00596-x

[65] M. E. Walton, S. W. Kennerley, D. M. Bannerman, P. E. M. Phillips, and M. F. S. Rushworth. 2006. Weighing up the benefits of work: Behavioral and neural analyses of effort-related decision making. *Neural Networks* 19, 8 (2006), 1302–1314. DOI: https://doi.org/10.1016/j.neunet.2006.03.005 Neurobiology of Decision Making.

Received 26 May 2023; revised 6 September 2023; accepted 15 October 2023