Are we measuring trust correctly in explainability, interpretability, and transparency research?

TIM MILLER, School of Computing and Information Systems and the Centre for AI & Digital Ethics, The University of Melbourne, Australia

This paper presents an argument for why we are not measuring trust sufficiently in explainability, interpretability, and transparency research. Most studies ask participants to complete a trust scale to rate their trust of a model that has been explained/interpreted. If the trust is increased, we consider this a positive. However, there are two issues with this. First, we usually have no way of knowing whether participants should trust the model. Trust should surely decrease if a model is of poor quality. Second, these scales measure perceived trust rather than demonstrated trust. This paper showcases three methods that do a good job at measuring perceived and demonstrated trust. It is intended to be starting point for discussion on this topic, rather than to be the final say. The author invites critique and discussion.

Additional Key Words and Phrases: trust, reliance, explainability, interpretability, transparency

ACM Reference Format:
Tim Miller. 2022. Are we measuring trust correctly in explainability, interpretability, and transparency research?. In *TRAIT 2022: Trust and Reliance in AI-Human Teams.* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As Hoffman [9] notes, the idea of asking whether someone trusted their computer was once considered a strange question not worthy of asking. Why would people trust computers? As we have delegated more tasks to machines, trust in these machines has become of interest in areas from computer science to human factors. Parasuraman and Riley's [23] seminal article on the use, misuse, disuse, and abuse of automation foreshadowed the problems of over- and under-reliance on machines due to lack of trust, while Lee and See [18] gave perhaps the first conceptual model of trust in machines and its processes that helped to clarify trust, reliance, and machines for many researchers.

In most research on trust and machines, trust is considered a mental attitude, such as a belief. The general hypothesis is that in most contexts, if a person trusts a machine[1], they will be more likely to rely on it. If they do not trust a machine, they will be less likely to rely on it, perhaps rejecting it entirely. Problems result when the alignment between whether a person trusts (distrusts) a machine does not align with whether they *should* trust (distrust) a machine. As Parasuraman and Riley [23] show, both under-reliance and over-reliance on machines can be problematic, causing issues such as physical, mental, or economic harm. One aim when building products and services is to engender **properly calibrated trust**: people trusting the parts that are trustworthy and distrusting the parts that are not trustworthy.

[1] Throughout this paper, the term 'machine' is used a general term to describe computers, their software, individual applications, or even individual functions within applications.

Over this period, researchers have also struggled with how to **measure trust**. Given that it is a mental attitude, this must be measured in field studies, lab experiments, and via surveys/interviews, with human participants. Much of this research has resulted in outputs like scales and surveys that ask study participants to rate various attributes of trust.

However, these approaches measure **perceived trust**. While the perception of trust is an important thing to measure (it affects appropriation, adoption, and reliance), having participants merely state their trust is not the same as **demonstrating trust**. As such, other research has looked at how to measure trust via demonstration, such as the *trust fall game* [21], in which participants need to decide whether to use a particular agent to act on their behalf, demonstrating trust if they do. Further, even those these methods measure perceived trust, they are not designed to measure the effect of explainability, interpretability, and transparency (XIT) methods on the trust of participants.

In this paper, we look at methods to measure the **effect** that a XIT method has on trust. The key aspect that makes measuring this different to existing methods such as the trust fall game, is that we are not concerned with the trust of the underlying machine — we are instead concerned with the effect of using different XIT methods, which we call *interventions*. We discuss how to measure the effect of XIT methods on both calibrated trust, both perceived and demonstrated.

The key insight is that to measure whether XIT methods have an effect on calibrated trust, we must know whether the participant should consider underlying model being explained/interpreted is trustworthy. As this is not possible to know *a priori*, trust evaluation methods should **manipulate** the trustworthiness of models and then measure whether the XIT methods are able calibrate trust (distrust) for more (less) trustworthy systems.

The trust evaluation methods in this paper are not new – they are taken from existing literature. This paper aims to: (a) make the problems of measuring trust for XIT methods better understood by the research community; (b) share trust evaluation methods with this community; and (c) start a discussion on other ways to approach the problem.

2 RELATED WORK

In this section, we give a high-level overview of existing research on trust measurement.

2.1 Measuring Perceived Trust

Measuring perceived trust has received more attention than measuring demonstrated trust. Much of the work has focused on defining what constitutes trust to enough fine grain that we can measure the components of trust.

Jian et al. [15] are the first researchers to empirically derive a questionnaire for human-human trust and human-machine trust. The questionnaire is mostly targeted towards automation/autonomy, given the terms used. Jian et al. propose a trust checklist consisting of 12 Likert-scale based items, ranging from deceit, to reliability, to integrity.

Since then, other scales of trust have been proposed, often with specific use cases in mind; e.g. Cahour and Forzy [4], Wang et al. [27], Wang and Moulden [26]. These scales are typically based on Jian et al.'s [15] original scale, and there is high overlap in the scale components, although Wang and Moulden [26] propose a scale specific to data-driven AI methods. Dizaji and Hu [6] conduct a comprehensive review of trust measurement scales, viewing them through the lens of the IMPACTS model of trust [11]. Their main finding is that the component of *adaptivity* is not well considered in trust and automation research.

Most relevant to this paper is the trust scale developed by Hoffman et al. [10], which is a trust scale aimed at explainable AI. Like other scales, it is based on Jian et al. [15], with some items adapted from Cahour and Forzy [4] and Wang et al. [27]. However, while for use in explainable AI, the scale proposed by Hoffman [9], does not explicitly measure the effect of an XIT intervention, as it is a scale, not a process. This scale and others outlined in this section can be used in a larger evaluation process to measure the effect of XIT methods, as we outline in Section 5.

2.2 Measuring Demonstrated Trust

Measuring demonstrated trust has also received some attention in the literature. Later in Section 5 we discuss work from Hussein et al. [13], Huber et al. [12] and Schmidt and Biessmann [24], who propose and/or use measures of demonstrated trust.

The idea of demonstrated trust between people is seen in areas such as supply chains [17, 25], organisations [20], journalism [7], and online interactions [22]. But perhaps the most explicit research that looks at demonstrated trust is in economic game theory. The *investment game* [2, 5] is a game in which participants are given real amounts of money and must 'invest' it. The payoffs of the players are dependent on each other, not just their own actions. By varying the rules and amounts, researchers can determine the amount of trust between players. However, this work measures interpersonal trust between two or more people, rather than the uni-directional trust between a human and a machine.

The *trust fall game*, proposed by Miller et al. [21], aims to measure demonstrated trust in machines using reliance. That is, trust is demonstrated by relying on an agent. The game is based on the well-known 'trust fall' activity, in which a person, the *trustor* closes their eyes, falls backwards, and one or more people, the *trustees*, catch them. Anyone who does not trust the people catching them will not submit themselves to this test. Miller et al. [21] use this analogy for a trust fall game for machines. Participants in an evaluation first establish a mental model by observing or interacting with a machine. Then, participants' trust is then tested by giving them to choice to rely on the machine's answers to problem, or to solve it themselves. The hypothesis is that if the participant has deemed the machine to be trustworthy, they are more likely to rely on the machine than if they deem the machine to not be trustworthy.

The trust fall game is a nice model for evaluating **demonstrated trust**, however, on its own, it is not sufficient for measuring the effect of an intervention such as a XIT method. As there is no way to assert what level of trust a participant *should* trust (or distrust) the agent, it is not possible to say if the XIT intervention is working as intended.

3 TRUST IN HUMAN-MACHINE SCENARIOS

In this section, we present an existing definition of trust in human-machine scenarios, which we use as a basic for defining the parameters of good evaluation methods.

Definition 3.1 (Interpersonal trust Mayer et al. [19]). Mayer et al. [19] 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."

Jacovi et al. [14] build on this model by first acknowledging that human-machine trust is similar to directional, interpersonal trust, except where the trustee is a machine doing some action, such as providing advice, finding information, executing a command, etc. Jacovi et al.'s definition goes further to consider two additional key elements: (1) the *context* in which the action is to be performed; and (2) *what* exactly is the machine being trust with.

The *what* part in Jacovi et al.'s model insists that trust be considered specifically against a set of *contracts*, which can be a legal or social agreements that document the expected behaviour of a machine for given tasks. For example, there may be a social expectation that a machine learning model can accurately and fairly predict the right products to promote to people online, but not necessarily in an explainable manner. As another example, in an air control system, there may be a legal contract spelling out that calculations of collisions between aircraft must be correct, but also completed within a specific time limit.

Bringing Definition 3.1 and the notions of context and contract, Jacovi et al. define human-machine trust as follows.

Definition 3.2 (Human-machine trust and distrust [14]). If and only if a person believes that a machine M will uphold a contract C in context n, and the person accepts vulnerability to M's actions, then that person **trusts** M contractually to C in n.

If and only if a person does not accept vulnerability to machine M's actions *because* they believe that M will fail to uphold a contract C in context n, then that person **distrusts** M to uphold contract C.

Definition 3.3 (Machine trustworthiness [14]). A machine is **trustworthy** to a contract C if and only if it is capable of maintaining this contract.

Often, we see the conflation of trustworthiness with trust, or definitions of trust that are defined in terms of trustworthiness. Jacovi et al.'s definition above cleanly separates trust and trustworthiness. This allows us to then define the model in Table 1. Inappropriately trusting when a something is not trustworthy is *unwarranted* trust; and a similar concept exists for *unwarranted distrust.* However, they mean that trust (and distrust) are **caused by** the ability to uphold a contract; or more succinctly, caused by the trustworthiness of the model, rather than being coincidental.

	Trustworthy	**Not trustworthy**
Trusted	Warranted trust	Unwarranted trust
Distrusted	Unwarranted distrust	Warranted distrust

Table 1. Warranted and unwarranted trust

From this model of warranted/unwarranted trust/distrust, we can clearly state that the goal with respect to trust should be **appropriately calibrated trust**. That is, we should aim to avoid unwarranted trust and unwarranted distrust, rather than just aim for trust irrespective of whether a machine is trustworthy for particular contracts/contexts.

The **four important lessons** from this section as related to evaluating trust are:

(1) the **vulnerability** of the trustee to the trustor, which implies that the trustee is taking a risk by willingly allowing themselves to be vulnerable;
(2) the **expectation** of the trustor that the trustee will perform the action in the best interests of the trustee;
(3) the outcome of that action is important to the trustor – they have a **stake** in the outcome; and
(4) warranted trust and distrust are caused by the trustworthiness of the trustee, and we should aim to **avoid unwarranted trust** as well as unwarranted distrust.

One may argue that this is perhaps an over-simplified model of trust, and they would be right that it fails to capture much of the realities of trust in human societies and human-machine trust. In reality, any type of trust is tentative and trust in machines is a dynamic process that changes over time as people interact with the machine. Further, it is not a binary concept, but a sliding scale, perhaps better represented by a probability or confidence. Nonetheless, we believe that this model is useful. We argue in the rest of this paper that in XIT research, we are failing to measure even such basic notions of trust.

4 REQUIREMENTS FOR EVALUATING WARRANTED/UNWARRANTED TRUST/DISTRUST

If we want to evaluate trust in field studies or lab studies that asses XIT methods, we argue that any evaluation must have the following requirements:

Task performance The tasks done by participants in the evaluation, whether part of an experiment or prior to any discussion, must have some measurable performance. That is, participants must be completing some task that has an outcome. **Rationale**: The goal of trust is not simply to have trust, but to support predictability in social interaction.

Risk Subjects must be vulnerable to the risk of the tasks being evaluated. That is, there must be some downside to having unwarranted trust or unwarranted distrust in the evaluation. Further, the participants must be aware of the risk. **Rationale**: Trust cannot exist without vulnerability. Kee and Knox [16] similarly argue that to study trust, any interaction must have risk, and the the participants in that interaction must be aware of them.

Further, to measure demonstrated trust, as opposed to (or in addition to) perceived trust, we have another requirement:

Reliance Participants must be given an opportunity to choose whether to rely on the machine. **Rationale**: For trust to exist, people must *accept* vulnerability to risk. Without having to rely on the machine in some way, there is no acceptance of the risk involved, and so no trust.

Methods for explainability, interpretability, and transparency are interventions that aim to manipulate (usually positively) people's understanding of models. To evaluate trust of **interventions**, we have one further requirement:

Manipulated trustworthiness There must be some known or estimated 'level' of trustworthiness that is manipulated as part of the evaluation. Trust measures must be taken for these different levels. **Rationale**: If we want to measure *warranted* trust (that is, caused by the trustworthiness), we cannot establish whether the intervention has correctly calibrated trust without manipulating the trustworthiness of the machine. For example, if we try technique A against baseline B, and technique A is shown to engender higher trust, this is meaningless if we do not know the trustworthiness. A good intervention should *decrease* trust if the model is less trustworthy than the trustee initially believes. Given that participants in studies would have little-to-no evidence to establish trust initially, manipulating trustworthiness to different, known/estimated levels allows a before/after comparison.

5 MEASURING TRUST

In this section, we showcase three methods for measuring the effect of trust on XIT interventions. All of these methods are extension to the idea of the *trust fall game* [21], discussed in Section 2.

5.1 Approaches to Measuring Trust

There are three broad categories of evaluation for measuring trust:

Questionnaires or surveys Participants in a study have access to some evidence of trustworthiness, and are asked to rate their opinion of trustworthiness. Hoffman et al. [10] present a survey instrument specifically for measuring trust within explainability (which in our view extends to any intervention intended to calibrate trust). It is important to emphasise: these measure participants' *perception* of their trust. It is well known that self reporting in experiments can be inaccurate, and even in the are of trust, experiments show that people's reported ratings of trust do not correspond with their actions [21].

- **Interviews, focus groups, and similar reporting tools** Researchers interview people to ask about their experiences and how this influences trust. These offer much richer insights into the mental model of participants, but are more labour intensive, and like questionnaires and surveys, they give insight into perceived trust, rather than actual trust.
- **Reliance** Participants are given the option to use or rely on something. The more often they choose not to rely on this, the less they are considered to trust it compared to the alternative. This demonstrates trust, rather than rating the perception of trust.

None of these methods is ‘the correct’ or best way to measure trust. While reliance is a better measure of calibrated trust, the perception of trust is also important. One would not want to design an explainability technique the increases reliance, but the perception of users is that they do not trust it. Ideally, we want perceived trust to align with reliance; and ideally, we would take measures using all three. Whichever category a particular measurement tool falls into, the requirements outlined in Section 4 must be considered if we want to gain proper insight into trust.

5.2 Measuring the Effect of XIT Interventions on Trust

In this section, we outline three protocols that can be used to evaluate trust in research where we are testing the effect of a XIT intervention on trust. These three protocols have been used by other researchers, and we identify papers where we first read about them.

5.2.1 Within-subject design. Figure 1 shows a within-subject design used by Huber et al. [12] in their experiments measuring the effects of using saliency maps and summaries for explaining agent policies.

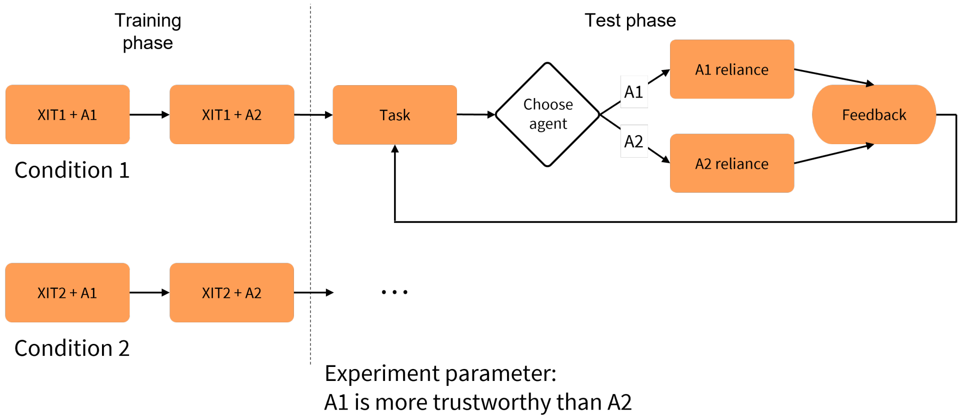

Fig. 1. A within-subject measure of demonstrated trust for XIT methods, used by Huber et al. [12]. Participants interact with two or more agents with differing levels of trustworthiness, and then must choose one to rely on to complete each task in a sequence of tasks. One condition for each method.

In this design, each participant acquires some experience of two or more agents of differing levels of trustworthiness. This could be a training phase of an experiment or practical experience in the field. The differing levels of trustworthiness are known by the researchers, but not the participants. This can be achieved by e.g. inserting random noise to reduce the performance of a machine learning classifier. Participants also interact with an XIT technique (or baseline) to help determine the trustworthiness of the agents. To avoid learning biases, the order in which the participant interacts with the agents must be counter-balanced.

In the test phase, participants must then choose between the two agents to do a task or series of tasks. Importantly, the output of the agent and the XIT technique is not provided during this test phase. Participants must make a judgement whether to rely on the agent without seeing its answer. This reliance demonstrates trust.

If there are multiple tasks in the test phase (e.g. multiple scenarios in a focus group), averaging the number of times each agent was used allows us to then give an overall ‘score’ for each agent. Given that agent A1 is more trustworthy than agent A2, the assumption of this design is that if an XIT technique is better for engendering warranted trust and distrust, then participants should be able to better determine the trustworthiness of agent A1. They would therefore rely on agent A1 more often than agent A2. Finally, a trust scale or survey can be administered to measure perceived trust.

One can see that this design adheres to our requirements from Section 4. First, there is a task that is completed throughout the study. Second, the participants are vulnerable to the risk of choosing the less trustworthy agent, provided this is part of the study (see Section 5.3 for more details on this). Third, the participants are given the choice to rely on one of the agents. Finally, the trustworthiness is manipulated by the researchers.

5.2.2 Between-subject design. Figure 2 shows a between-subject design for measuring trust. The rationale for this design is the same as the the within-subject design. The basic design is similar to the within-subject design from Section 5.2.1, with ‘known’ trustworthiness of two agents A1 and A2, and a choice. The difference is in two places:

(1) Each participant sees just one agent instead of two.
(2) The choice is between the agent and the participant performing the task themselves (self reliance).

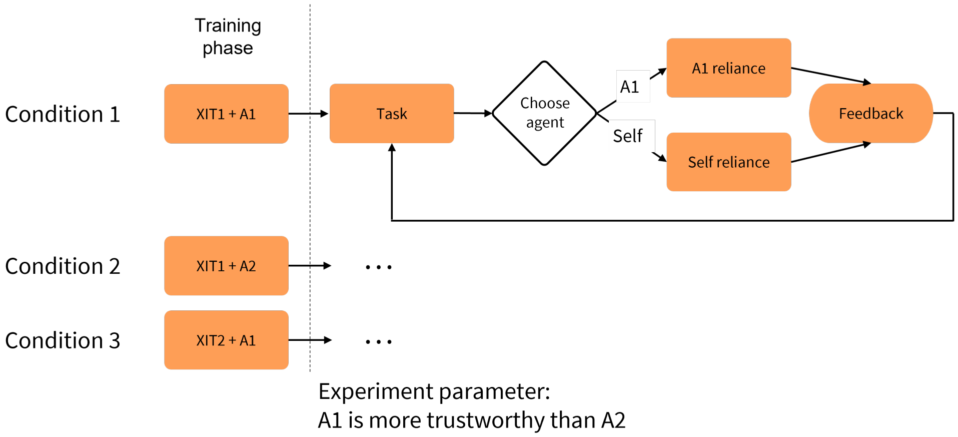

Fig. 2. A between-subject measure of demonstrated trust for XIT methods. Participants interact with an agent, and then must choose whether to rely on themselves or the agent to complete each task in a sequence of tasks. One condition for each pair in method × agent.

Hussein et al. [13] use a cross between these two designs in their experiments measuring the effects of transparency on human-agent decision making in swarm systems. Each participant is tasked with two of the conditions from Figure 2: one for each agent to account for trustworthiness, with the order counterbalanced to avoid ordering effects. The choice is between the agent and the participant, rather than the two agents. The researchers also task participants to fill out a (perceived) trust questionnaire at intervals.

The advantage of Hussein et al.'s design is that the researcher can perform mediation analysis on the relationship *Trustworthiness* ⟶ *Perceived Trust* ⟶ *Reliance* to determine how much the trustworthiness affects perceived trust, which in turn affects reliance, versus how much the trustworthiness affects reliance (or actual trust) on its own.

Within- vs. between-subject design. The main advantage of a between-subject design is any ordering effects are eliminated, rather than controlled for. However, a small advantage of a within-subject design, such as those used by Huber et al. [12] and Hussein et al. [13] is that we can run post-experiment surveys or interviews where the participants can give their impressions on the **difference** between the trustworthiness of the agents. This direct comparison gives participants reference points against which to compare agents.

5.2.3 Manipulating trust at the instance level. Schmidt and Biessmann [24] propose a different design in which the manipulation of trust is at the task or instance level. Rather than having different agents with different levels of trustworthiness, individual tasks or instances are chosen to be 'high' or 'low' quality. The overall framework is outlined in Figure 3. Instead of two or more agents, a single agent is used. The accuracy/performance of agent decisions in individual tasks is known; that is, there is a series of tasks/scenarios, and the ground truth of these is known. A trust score is calculated based on agreement between participant and agent. An important distinction between this design and the earlier two designs is that the participants see the 'explanation' (or XIT method) instead. This hypothesis is that a better XIT method will give a better understanding of the domain.

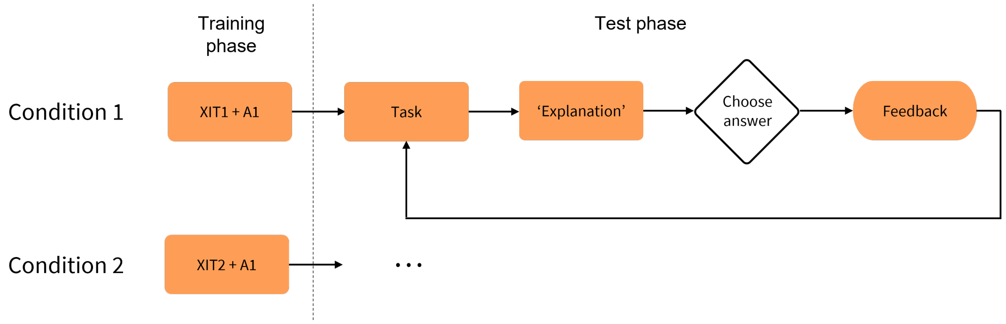

Fig. 3. A measure of trust for XIT methods based on information gain. Participants interact with an agent, and then must choose whether to rely on themselves or the agent to complete each task in a sequence of tasks. One condition for each pair in method × agent.

Schmidt and Biessmann [24] first define the *information transfer rate* (ITR) from an XIT method as:

$$ITR = \frac{I(Y_h,Y_a)}{t} \quad \text{st.} \quad I(Y_h, Y_h) = \Sigma_{y_a,y_h} p(y_a, y_h) \log \frac{p(y_a,y_h)}{p(y_a)p(y_h)}$$

where Y_h and Y_a are the set of answer given by the human (h) and agent (a) respectively, $y_h \in Y_h$ and $y_a \in Y_a$ are individual answers respectively, and t is the amount of time spent by the human participant answering the set Y_h. So, Schmidt and Biessmann define ITR using *information gain*. The more the human participant and the agent agree on answers, the higher the ITR. Time is used with the assumption that a better understanding provided by a XIT technique will enable faster answers. For the rest of this paper, we will not discuss t.

Trust is then defined as $T = \frac{ITR_{Y_a}}{ITR_Y}$, where ITS_{Y_a} is the ITR between the agent and the human participant, and ITR_Y is the ITR between the participant and the ground truth. So, trust is a fraction of the ITR of the agent+human over the ITR of the human+ground truth. A score greater than 1 indicates a participant who has 'too much' unwarranted trust in the agent, while a score less then 1 indicates a participant who has 'too much' unwarranted distrust. A score of 1 is 'perfectly calibrated trust'.

However, this measures agreement rather than reliance, so is not truly measuring *demonstrated* trust. Nonetheless, this approach can be modified to measure demonstrated trust: showing the task, withholding the explanation (and agent decision), and then asking the participant to choose between the agent and self reliance, as is done in Section 5.2.2.

An advantage of manipulating trust at the instance level is that it gives us a finer-grained control to measure trust calibration. By choosing instances based on their alignment with the ground truth, we can measure the calibration by measuring the number of false positives and negatives.

A disadvantage of this approach is that it requires ground truth, which is often not available or not really 'ground' at all. The previous approaches that manipulate trustworthiness by using different agents is easy to achieve by simply taking one agent and adding noise to their answers, as is done by Hussein et al. [13].

5.3 Vulnerability in laboratory experiments

'Vulnerability to risk' is a key component of trust. The designs above did not consider risk explicitly. In a questionnaire/survey or field study of trust, participants must be (or have been) vulnerable to the risk of the agent failing. Reliance without risk does not require trust. In experimental situations, participants must also be have stakes in the situation. This is difficult because in many cases, participants know that after any experiment, they can return to their normal lives with little impact. In such cases, we need to provide incentives that simulate risk.

There are two ways we have seen that can introduce risk:

Payment bonuses As noted earlier, because a good design requires that participants undertake some measurable tasks, we can provided bonuses on top of the base payment to participants who excel. For example, a small bonus per task that is completed well, or a single bonus for reaching a particular milestone. In some cases, we may want to measure the trust impact on participants playing the role of *decision subjects* — that is, simulating situations in which a *decision maker* uses an algorithm to make a decision about a decision subject, such a loan decision being made about a borrower. In this case, decision subject has no control over the outcome. In these cases, decision subjects receive bonus payments depending on the behaviour of other participants. This is less straightforward from an ethical perspective, but is not uncommon in experiments of bi-directional trust (see Section 2; and is a further reason why risk bonuses must be on top of the base payment for participants.

The reader may question whether missing out on bonus payments can be considered risks; and it is true that we cannot compare the risk losing of a small financial bonus with the risk of e.g. undergoing an invasive medical treatment. However, Bradler et al. [3] show that financial bonuses for performance increased output in both a creative task and a simple task; and studies such as those by Amir et al. [1] show that small bonuses affects people's decisions in economic games, even when played in online platforms like Amazon Mechanical Turk, where participants are quite anonymous to each other. So, we argue that participants are presented with a risk and they are vulnerable to this risk. As a result, we are **measuring trust** when we use financial bonuses for performance. However, we should be careful to not over-interpret the results as saying that our models are trustworthy for high stakes decisions when the laboratory stakes are low.

Gamification Tasking participants to play games with measurable performance can increase people's engagement and avoid player fatigue. Guttman et al. [8] argue that this makes them an good proxy for measuring human-AI interactions, compared to other synthetic tasks that are not gamified, because they "allow us to study how framing affects human-AI interfaces in more realistic ways than laboratory experiments alone".

We argue that the risk of losing points in a gamification environment is a reasonable proxy for low stakes risk; enough to measure trust as reliance. For example, in a laboratory experiment, we used gamification to increase participant engagement. Outside of the laboratory, some participants created a leaderboard. Once we have a handful of names on the leaderboard, people started asking if they could do the experiment again for no payment, so they could work their way up the leaderboard.

As with financial bonuses, the risk is low, so we should be careful how we interpret results from games when the stakes are low.

Deception A third approach is to deceive participants into thinking that they are at risk. There are clear ethical concerns with such an approach, but if handles ethically, can be a powerful method.

6 DISCUSSION

It is important to note that the evaluation methods in this paper are not restricted to laboratory experiments: they generalise to field studies. In addition, the general design of intervening on trustworthiness is required by all three types of measuring trust outlined in Section 5.1: questionnaires/surveys, interviews/focus groups, and reliance.

All three evaluation methods have some intervention on trustworthiness — either on the agent itself or on individual decisions. Can these approaches be used in field studies where we cannot manipulate trustworthiness? We believe they cannot; however, this is not a weakness of the approaches – it is the reality of measuring the effect of XIT methods on trust. Without manipulating trustworthiness, we are not truly measuring trust.

Note that the idea of manipulating trust is not only important for demonstrated trust, but also for perceived trust. As with the approached outlined in this paper, if we want to measure whether our XIT method *correctly* impacts the perceived trust of participants, we need to manipulate the trustworthiness of the underlying agent.

One final point to make is that trust is a dynamic concept, and people's trust in machines varies over time [9], even in the presence of XIT methods. As XIT researchers, we need to start evaluating the impact of XIT methods in more longitudinal studies, measuring trust as well as other crucial factors. That is not to say that such longitudinal studies are easy – they are difficult. However, any such studies will be highly valuable to the community.

This position paper is not intended to be the end of the conversation. We encourage people to critique these methods, propose changes, and propose new methods. The primary reason for this paper is to promote discussion and make the research community aware of existing work on measuring the effect of XIT interventions on trust.

ACKNOWLEDGMENTS

Thanks to Alon Jacovi, Ana Marasovic, and Yoav Goldberg for their discussions about trust and trustworthiness.

This research is supported by Australian Research Council (ARC) Discovery Grant DP190103414: *Explanation in Artificial Intelligence: A Human-Centred Approach*.

REFERENCES

[1] Ofra Amir, David G Rand, and Ya'akov Kobi Gal. 2012. Economic games on the internet: the effect of $1 stakes. *PLoS One* 7, 2 (Feb. 2012), e31461. https://doi.org/10.1371/journal.pone.0031461

[2] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 1 (1995), 122–142.

[3] Christiane Bradler, Susanne Neckermann, and Arne Jonas Warnke. 2019. Incentivizing Creativity: A Large-Scale Experiment with Performance Bonuses and Gifts. *J. Labor Econ.* 37, 3 (July 2019), 793–851. https://doi.org/10.1086/702649

[4] Béatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? An example with a cruise control system. *Safety science* 47, 9 (2009), 1260–1270.

[5] Colin Camerer and Keith Weigelt. 1988. Experimental tests of a sequential equilibrium reputation model. *Econometrica: Journal of the Econometric Society* (1988), 1–36.

[6] Lida Ghaemi Dizaji and Yaoping Hu. 2021. Building And Measuring Trust In Human-Machine Systems. In *2021 IEEE International Conference on Autonomous Systems (ICAS)*. 1–5. https://doi.org/10.1109/ICAS49788.2021.9551131

[7] Katherine M Engelke, Valerie Hase, and Florian Wintterlin. 2019. On measuring trust and distrust in journalism: Reflection of the status quo and suggestions for the road ahead. *Journal of Trust Research* 9, 1 (Jan. 2019), 66–86. https://doi.org/10.1080/21515581.2019.1588741

[8] Rotem D Guttman, Jessica Hammer, Erik Harpstead, and Carol J Smith. 2021. Play for Real(ism) - Using Games to Predict Human-AI interactions in the Real World. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY (Oct. 2021), 1–17. https://doi.org/10.1145/3474655

[9] Robert R Hoffman. 2017. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive systems engineering: The future for a changing world* (2017), 137–164.

[10] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[11] Ming Hou, Geoffrey Ho, and David Dunwoody. 2021. IMPACTS: a trust model for human-autonomy teaming. *Hum.-Intell. Syst. Integr.* 3, 2 (June 2021), 79–97. https://doi.org/10.1007/s42454-020-00023-x

[12] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. 2021. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence* 301 (2021), 103571.

[13] Aya Hussein, Sondoss Elsawah, and Hussein A Abbass. 2020. Trust mediating reliability–reliance relationship in supervisory control of human–swarm interactions. *Human Factors* 62, 8 (2020), 1237–1248.

[14] Alon Jacovi, Ana Marasovic, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2021)*. https://arxiv.org/abs/2010.07487

[15] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *Int. J. Cogn. Ergon.* 4, 1 (March 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[16] Herbert W. Kee and Robert E. Knox. 1970. Conceptual and methodological considerations in the study of trust and suspicion1: Conceptual Considerations Methodological Considerations Summary and Conclusions REFERENCES. *The Journal of Conflict Resolution (pre-1986)* 14, 3 (09 1970), 357.

[17] Mohammed Laeequddin, B S Sahay, Vinita Sahay, and Waheed K Abdul. 2010. Measuring trust in supply chain partners' relationships. *Measuring Business Excellence* 14, 3 (Jan. 2010), 53–69. https://doi.org/10.1108/13683041011074218

[18] John D Lee and Katrina A See. 2004. Trust in automation: designing for appropriate reliance. *Hum. Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[19] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An Integrative Model Of Organizational Trust. *AMRO* 20, 3 (July 1995), 709–734. https://doi.org/10.5465/amr.1995.9508080335

[20] Bill McEvily and Marco Tortoriello. 2011. Measuring trust in organisational research: Review and recommendations. *Journal of Trust Research* 1, 1 (April 2011), 23–63. https://doi.org/10.1080/21515581.2011.552424

[21] David Miller, Mishel Johns, Brian Mok, Nikhil Gowda, David Sirkin, Key Lee, and Wendy Ju. 2016. Behavioral Measurement of Trust in Automation: The Trust Fall. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet.* 60, 1 (Sept. 2016), 1849–1853. https://doi.org/10.1177/1541931213601422

[22] Jason R C Nurse, Ioannis Agrafiotis, Michael Goldsmith, Sadie Creese, and Koen Lamberts. 2014. Two sides of the coin: measuring and communicating the trustworthiness of online information. *Journal of Trust Management* 1, 1 (May 2014), 1–20. https://doi.org/10.1186/2196-064X-1-5

[23] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum. Factors* 39, 2 (June 1997), 230–253. https://doi.org/10.1518/001872097778543886

[24] Philipp Schmidt and Felix Biessmann. 2019. Quantifying Interpretability and Trust in Machine Learning Systems. (Jan. 2019). arXiv:1901.08558 [cs.LG] http://arxiv.org/abs/1901.08558

[25] Gaurav Tejpal, R K Garg, and Anish Sachdeva. 2013. Trust among supply chain partners: a review. *Meas. Bus. Excel.* 17, 1 (March 2013), 51–71. https://doi.org/10.1108/13683041311311365

[26] Jennifer Wang and Angela Moulden. 2021. AI Trust Score: A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21, Article 54)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3443452

[27] Lu Wang, Greg A Jamieson, and Justin G Hollands. 2009. Trust and reliance on an automated combat identification system. *Human factors* 51, 3 (2009), 281–291.