Formalizing Trust in Artificial Intelligence:
Prerequisites, Causes and Goals of Human Trust in AI

Alon Jacovi
Bar Ilan University
alonjacovi@gmail.com

Ana Marasović
Allen Institute for Artificial Intelligence
University of Washington
anam@allenai.org

Tim Miller
School of Computing and Information Systems
The University of Melbourne
tmiller@unimelb.edu.au

Yoav Goldberg
Bar Ilan University
Allen Institute for Artificial Intelligence
yoav.goldberg@gmail.com

ABSTRACT
Trust is a central component of the interaction between people and AI, in that ‘incorrect’ levels of trust may cause misuse, abuse or disuse of the technology. But what, precisely, is the nature of trust in AI? What are the prerequisites and goals of the cognitive mechanism of trust, and how can we cause these prerequisites and goals, or assess whether they are being satisfied in a given interaction? This work aims to answer these questions. We discuss a model of trust inspired by, but not identical to, sociology’s interpersonal trust (i.e., trust between people). This model rests on two key properties of the vulnerability of the user and the ability to anticipate the impact of the AI model’s decisions. We incorporate a formalization of ‘contractual trust’, such that trust between a user and an AI is trust that some implicit or explicit contract will hold, and a formalization of ‘trustworthiness’ (which detaches from the notion of trustworthiness in sociology), and with it concepts of ‘warranted’ and ‘unwarranted’ trust. We then present the possible causes of warranted trust as intrinsic reasoning and extrinsic behavior, and discuss how to design trustworthy AI, how to evaluate whether trust has manifested, and whether it is warranted. Finally, we elucidate the connection between trust and XAI using our formalization.

CCS CONCEPTS
• Human-centered computing → HCI theory, concepts and models; • Applied computing → Sociology; Psychology; • Social and professional topics → Computing / technology policy; • Computing methodologies → Artificial intelligence; Machine learning.

KEYWORDS
artificial intelligence, trust, trustworthiness, machine learning, risk, vulnerability, sociology, explainability, interpretability

1 INTRODUCTION
With the rise of opaque and poorly-understood machine learning models in the field of AI, trust is often cited as a key desirable property of the interaction between any user and the AI [8, 11, 71, 76]. The recent rapid growth in the field of explainable AI (XAI) is also, in part, motivated by the need to maintain trust between the human user and the AI [31, 35, 46, 52, 59, 72]. By designing AI that users can and will trust to interact with, the AI can be safely implemented in society.

However, literature seldom discusses specific models of trust between humans and AI. What, precisely, are the prerequisites for trust between a user and AI? For what goals does the cognitive mechanism of trust exist? How can we design AI which facilitates these prerequisites and goals? And how can we assess whether the prerequisites exist, and whether the purpose behind the trust has been achieved?

In this work, we are interested in formalizing the ‘trust’ transaction between the user and AI, and using this formalization to further our understanding of the requirements behind AI which can be integrated in society. We consider ‘artificial intelligence’ to be any automation which is attributed with intent by the user [social attribution, 52], i.e., anthropomorphized with a human-like reasoning process. For our purpose, we consider the user to be an individual person, rather than an organization, though aspects of the work are applicable to the latter as well.

There are many vague and unclear aspects of trust which are difficult to formalize with the tools available to us in literature on AI and Human-Computer Interaction (HCI). For this reason, we first discuss how interpersonal trust is defined in sociology, and derive a basic, yet functional, definition of trust between a human and an AI, based on the prerequisites and goals behind the trustor developing trust in the AI (Section 2). Specifically, the trustor must be vulnerable to the AI’s actions, and the trustor’s goal in developing trust is to anticipate the impact of the AI model’s decisions.

However, the above definition is incomplete: though the goal is anticipating ‘intended’ behavior, what can we say about when and whether this goal is achieved? We develop the definition further by answering two questions: (1) what is the AI being trusted with (i.e., what is ‘intended’)? and (2) what differentiates trust which achieves this goal, and trust which does not? Section 3 answers (1) via a notion of contractual trust, and Section 4 answers (2) via notions...
of warranted and unwarranted trust. In Section 5 we complete the definition of Human-AI trust with a formal summary of the above.

With these definitions, we are now equipped to discuss the causes of trust in the AI (specifically, warranted trust in a particular contract), and how we should pursue the development of AI that will be trusted. In Section 6, we answer the question: what are the mechanisms by which an AI model gains the trust of a person? Namely, we define and formalize notions of intrinsic trust, which is based on the AI’s observable reasoning process, and extrinsic trust which is based on the AI’s external behavior.

Both intrinsic and extrinsic trust are deeply related to XAI. As mentioned, XAI literature frequently notes trust as a principal motivation in the development of explanations and interpretations in AI, but seldom elucidates the precise connection between the methods and the goal. In Section 7, we unravel this ‘goal’ of XAI—to facilitate trust—by utilizing our formulation thus far.

In Section 8 we pivot to the question of evaluating trust, by discussing the evaluation of the vulnerability in the interaction and the evaluation of the ability to anticipate. Finally, in Section 9 we expand on other aspects of interpersonal trust and general human-machine trust (i.e., automation which is not attributed with intent), how they relate to our notion of Human-AI trust, and possible future extensions of our formalization.

Contributions. We provide a formal perspective of Human-AI trust which is rooted in, but nevertheless not the same as, sociology’s perspective of interpersonal trust. We use this formalization to inform notions of the causes behind Human-AI trust, the connection between trust and XAI, and the evaluation of trust. We hope that this work enables a principled approach to developing AI that should, and will, be trusted in practice.

Note on the organization of the work. The following sections provide an informal description of trust in AI via a narrative, in the interest of accessibility (§2,3,4). We provide formal, concise definitions of our taxonomy after completing the relevant explanations (§5). Additionally, for coherency we bypass some nuanced reasoning behind our choice of formalization, which we make available at the end of the work (§9).

2 A BASIC DEFINITION OF TRUST

To understand human trust in AI (Human-AI trust), a useful place to start is to examine research in philosophy, psychology, and sociology of how people trust each other (interpersonal trust). In this section, we present a primitive (and incomplete, as we will show) definition of trust that will serve as a basis for the rest of the work.

Definition (interpersonal trust). A common basic definition of trust regards it as a directional transaction between two parties: if A believes that B will act in A’s best interest, and accepts vulnerability to B’s actions, then A trusts B [49]. The goal of trust is to “make social life predictable [by anticipating the impact of behavior], and make it easier to collaborate between people” [53].

Noteworthy in this definition, and key to defining Human-AI trust, are the notions of anticipation and vulnerability. In particular, interpersonal trust exists to mitigate uncertainty and risk of collaboration by enabling the trustee’s ability to anticipate the trustee—where ‘anticipating’ refers to a belief that the trustee will act in the trustee’s best interests. We maintain that Human-AI trust exists for the same purpose, as a sub-case of trust in automation, following Hoffman [28]: trust is an attempt to anticipate the impact of behavior under risk. Based on this, we conclude:

Risk is a prerequisite to the existence of Human-AI trust. We refer to risk as a disadvantageous or otherwise undesirable event to the trustor (that is a result of interacting with the trustee), which can possibly—but not certainly—occur [24]. Therefore, "to act in A’s best interest" is to avoid the unfavorable event. Admitting vulnerability means that the trustor perceives both of the following: (1) that the event is undesirable; and (2) that it is possible. Ideally, the existence of trust can only be verified after verifying the existence of risk, i.e., by proving that both conditions hold.

For example, Al-produced credit scoring [9] represents a risk to the loan officer: a wrong decision carries a risk (among others) that the applicant defaults in the future. The loss event must be undesirable to the user (the loan officer), who must understand that the decision (credit score) could theoretically be incorrect, and that it is not certainly incorrect, for trust to manifest. Similarly, from the side of the applicants (if they have a choice as to whether to use the AI), the associated risk is to be denied (or to be charged a higher interest rate) on a loan that they deserve, and trust manifests if they believe that the AI will work in their interest (the risk will not occur).

Distrust manifests in attempt to mitigate the risk. The notion of distrust is important, as it is the mechanism by which the user attempts to avoid the unfavorable outcome. We adapt Tallant’s definition of distrust: A distrusts B if A does not accept vulnerability to B’s actions, because A believes that B may not act in A’s best interest [67]. Importantly, distrust is not equivalent to the absence of trust [50], as the former includes some belief, where the latter is lack of belief—or in other words, distrust is trust in the negative scenario. For the remainder of this paper, we focus our analysis on trust, as the link to distrust is straightforward.

The ability to anticipate is a goal, but not necessarily a symptom of Human-AI trust. The ability or inability of the user to anticipate the behavior of an AI model in the presence of uncertainty or risk, is not indicative of the existence or absence of trust. We illustrate this in §4. We stress that anticipating intended behavior is the user’s goal in developing trust, but not necessarily the AI developer’s goal.

3 CONTRACTUAL TRUST

The above notion of anticipating ability is incomplete. If the goal of trust is to enable the trustor’s ability to anticipate, what does the human trustor anticipate in the AI model’s behavior? And what is the role of the ‘anticipated behavior’ in the definition of Human-AI trust?
3.1 Trust in Model Correctness

XAI research commonly refers to the trust that the model is correct [e.g., 22, 46, 60]. What does this mean, exactly?

To illustrate this question, consider some binary classification task, and suppose we have a baseline that is completely random by design, and a trained model that achieves the performance of the random baseline (i.e., 50% accuracy in this case). Since the trained model performs poorly, a simple conclusion to draw is that we cannot trust this model to be correct. But is this true?

Suppose now that the trained model with random baseline performance does not behave randomly. Instead, it is biased in a specific manner, and this bias can be revealed with an interpretation or explanation of the model behavior. This explanation reveals to the user that on some types of samples, the model—which maintains random baseline performance—is more likely to be correct than for others. As an illustrative example, consider a credit-scoring AI model which is more likely to be correct for certain sub-populations.

The performance of the second model did not change, yet we can say that now, with the added explanation, a trustee may have more trust that the model is correct (on specific instances). What has changed? The addition of the explanation enabled the model to be more predictable, such that the user can now better anticipate whether the model’s decision is correct or not for given inputs (e.g., by looking at whether the individual is part of a certain sub-population), compared to the model without any explanation. Note that this is merely refers to one ‘instance’ of anticipation; it refers to anticipating a particular attribute of the AI’s decision (correctness), whereas in the previous definition (§2), it refers to general behavior.

We arrive at a more nuanced and accurate view of what “trust in model correctness” refers to: it is in fact not trust in the general performance ability of the model, but that the patterns that distinguish the model’s correct and incorrect cases are available to the user.

3.2 The General Case: Trust in a Contract

The above example of model correctness is merely an instance of what Hawley [26], Tallant [67] refer to as trust with commitment or contractual trust. Contractual trust is a model of trust in which a trustee has a belief that the trustee will stick to a specific contract.

In this work, we contend that all Human-AI trust is contractual, and that regardless of what the contract in is a particular interaction, to discuss Human-AI trust, the contract must be explicit.

Generally, the contract may refer to any functional which is deemed useful, even if it is not concrete performance at the end-task that the model was trained for. Therefore, model correctness is only one instance of contractual trust. For example, a model trained to classify medical samples into classes can reveal strong correlations between attributes for one of those classes, giving leads to research on causation between them, even if the model was not useful for the original classification task [42, 46].

Assume that the performance evaluation is representative of real usage for now, although this is an important factor that we will discuss in Section 6.2.

E.g., calibrated probabilities [45], where the classification probabilities of a model are calibrated with some measure of its uncertainty, can produce this effect.

To our knowledge Hawley [26] is the first to formalize trust as ‘trust with commitment’ as trust with commitment [e contract]. Tallant [67] expands on their work with terminology of contractual trust.

Although they do not refer to contractual trust in their work, Hoffman [28] provide support to formalize trust in automation (beyond AI) as multi-dimensional (which we interpret as multi-contractual), rather than a binary variable or sliding scale.

Contracts and contexts. The idea of context is important in trust: people can trust something in one context but not another [28]. For example, a model trained to classify medical samples into classes can perform strongly for samples that are similar to those in its training set, but poorly on those where some features were infrequent, even though the ‘contract’ appears the same. Therefore, contractual trust can be stated as being conditioned on context. For readability in the rest of this paper, we omit context from the discussion, but implicitly, we consider the contract to be conditioned on, and thus include, the context of the interaction.

What are useful contracts? The European Commission has outlined detailed guidelines on what should be required from AI models for them to be trustworthy (see Table 1, col. 1–2). Each of these requirements can be used to specify a useful contract.

Another area of research that is relevant for defining contracts is the work that proposes standardized documentations to communicate the performance characteristics of trained AI models. The examples of such documentations are: data statements [6], datasheets for datasets [18], model cards [54], reproducibility checklists [56], fairness checklists [48], and factsheets [2].

We illustrate the connection between these documentations and the European requirements in Table 1. For example, if transparency is the stated contract then all of the mentioned documentations could be used to specify information that AI developers need to provide such that they can evaluate and increase users’ trust in transparency of an AI system.

Explanation and analysis types depend on the contract. We argue that “broad trust” is built on many contracts, each involving many factors and requiring different evaluation methods. For example, the models’ efficiency in terms of the number of individual neurons responsible for a prediction is relevant for sustainability, but likely not for, e.g., ensuring accessibility.

We have previously illustrated that the addition of explanation of the model’s behavior can increase users’ trust based on one contract (§3.1). Just as different evaluation methods are needed for different types of contractual trust, so are different types of explanations. In Table 1, we outline different established types of explanatory methods and analyses that could be suitable for increasing different types of contractual trust derived from the European requirements.

Conclusions. The formalization of contracts allows us to clarify the goal of anticipation in Human-AI trust: contracts specify the behavior to be anticipated, and to trust the AI is to believe that a set of contracts will be upheld.

Specific contracts have been outlined and explored in the past when discussing the integration of AI models in society. We advocate for adoption of the taxonomy of contracts in Human-AI trust; for three reasons: (1) it has, though recent, precedence in sociology; (2) it opens a general view of trust as a multi-dimensional transaction, for which all relevant dimensions should be explored before integration in society; and importantly, (3) the term implies an obligation by the AI developer to carry out a prior agreement.

The guidelines are available at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
Table 1: The European requirements for trustworthy AI, related available documentation, and related explanatory methods or analyses. We position the European guidelines as a guidance on which contracts (§3) it is useful to pursue trust.

| European Guidelines for Trustworthy AI Models | Key Requirements | Factors | Documentations | Explanatory Methods/Analyses |
|---------------------------------------------|-----------------|---------|----------------|----------------------------|
| Human agency and oversight                   | Foster fundamental human rights | Fairness checklists | See "Diversity, non-discrimination, fairness" |
|                                             | Support users’ agency | All | User-centered explanations [59] |
|                                             | Enable human oversight | - | Explanations in recommender systems [41] |
| Technical robustness and safety              | Resilience to attack and security | Factsheets (security) | Adversarial attacks and defenses [20] |
|                                             | Fallback plan and general safety | Model cards (metrics) | - |
|                                             | A high level of accuracy | Factsheets (concept drift) | - |
|                                             | Reliability | Reproducibility checklists | Contrast sets [17], behavioral testing [58] |
|                                             | Reproducibility | - | “Show your work” [14] |
| Privacy and data governance                  | Ensure privacy and data protection | Datasheets/statements | Detection of protected attributes [57] |
|                                             | Ensure quality and integrity of data | Datasheets/statements | - |
|                                             | Establish data access protocols | Datasheets/statements | - |
| Transparency                                 | High-standard documentation | All | Saliency maps [61], self-attention patterns [40], influence functions [38], probing [16] |
|                                             | Technical explainability | Factsheets (explainability) | Counterfactual [21], contrastive [51], natural language [27], by-example [38], and concept-level [19] explanations |
|                                             | Adaptable user-centered explainability | Factsheets (explainability) | - |
|                                             | Make AI systems identifiable as non-human | - | - |
| Diversity, non-discrimination, fairness      | Avoid unfair bias | Fairness checklists | Debiasing using data manipulation [66] |
|                                             | Encourage accessibility and universal design | - | - |
|                                             | Solicit regular feedback from stakeholders | Fairness checklists | - |
| Societal and environmental well-being        | Encourage sustainable and eco-friendly AI | Reproducibility checklists | Analayzing individual neurons [10] |
|                                             | Assess the impact on individuals | Fairness checklists | Bias exposure [65] |
|                                             | Assess the impact on society and democracy | Fairness checklists | Explanations designed for applications such as fact checking [3] or fake news detection [47] |
| Accountability                               | Audibility of algorithms/data/design | Factsheets (lineage) | - |
|                                             | Minimize and report negative impacts | Fairness checklists | - |
|                                             | Acknowledge and evaluate trade-offs | - | Reporting the robustness-accuracy [1] trade-off or simplicity-equity trade-off [37] |
|                                             | Ensure redress | Fairness checklists | - |

4 TRUSTWORTHY AI

Trust, as mentioned, aims to enable the ability to anticipate intended behavior through the belief that a contract will be upheld. Further, as mentioned in Section 2, the ability to anticipate does not necessarily manifest with the existence of trust; it is possible for a user to trust a model despite their inability to anticipate its behavior. In other words, the belief exists, but may or may not be followed by the desired behavior. What differentiates trust that ‘succeeds’ at this goal from trust that does not?

Let us separate the two cases from the perspective of AI: given that the user trusts the AI model, anticipation depends on whether the model is able to carry out its contract. This perspective distinguishes “trust” (an attitude of the trustor) from being “trustworthy” (a property of the trustee) [64, 75], and we say that an AI model is trustworthy to some contract if it is capable of maintaining this contract.

Trust and trustworthiness are entirely disentangled: pursuing one does not entail pursuing the other, and trustworthiness is not a prerequisite for trust, in that trust can exist in a model which is not trustworthy, and a trustworthy model does not necessarily gain trust. We say that the trust is warranted if it is the result of trustworthiness, and otherwise it is unwarranted [50]. Warranted trust is sometimes referred to as trust which is calibrated with trustworthiness [43]. In other words, trust is the cognitive mechanism to give the ‘illusion’ of anticipating intended behavior, which becomes reality when the trust is warranted, and the trustor feels “betrayed” when the illusion is broken.

For example, consider a user interacting with an AI model via some visual interface (GUI), and the user trusts the AI to make a correct prediction on some task. There is a correlative, but not causal, relationship between high-quality GUI and trustworthy AI models due to a shared variable of high budget in the development of the AI. If the cause of the user’s trust is the model GUI, then manipulation of the model’s ability to make good predictions will not affect this trust, and thus it is unwarranted. If the cause of the trust is the model’s performance ability (due to its higher budget), then theoretically manipulating this performance level will affect the level of trust. This example is illustrated in Figure 1.
Formally, we define warranted Human-AI trust via a causal (interventionist) relationship with trustworthiness: incurred Human-AI trust is warranted if the trustworthiness of the model can be theoretically manipulated to affect the incurred trust. Note that by this definition, it is possible for a trustworthy model to incur unwarranted Human-AI trust—in this case, the trust will not be betrayed, even though it is unwarranted.

When XAI literature refers to trust, we assume that it is referring to trust which is warranted. Therefore, we contend that when pursuing Human-AI trust, unwarranted trust should be explicitly evaluated against, and avoided or otherwise minimized. See [55] for analysis on the dangers of human usage (or avoidance) of automation when trust and trustworthiness are misaligned, on axes of disuse, misuse and abuse. Specifically, trust exceeding trustworthiness leads to misuse, while trustworthiness exceeding trust leads to disuse.

Finally, the notion of warranted distrust is similar to that of warranted trust, and easily derived from it: we say that the distrust is warranted if it is sourced in the non-trustworthiness of the AI, i.e., the lack of capability of the AI to maintain the contract, and otherwise, it is unwarranted. It stands to ethics that if the AI is incapable of maintaining some relevant contract, it is a desired outcome (desired by the AI developer) that the user develop warranted distrust in that contract, which will be beneficial in applying the AI to some scenario in spite of its flaws.

---

**5 DEFINING HUMAN-AI TRUST**

This section serves as a formal definition of the taxonomy thus far. Note that we consider AI as an automation which is attributed with human-like intelligence by the human interacting with it.\(^8\)

**Trustworthy AI.** An AI is trustworthy to contract C if it is capable of maintaining the contract.

**Human-AI trust.** If H (human) perceives that M (AI model) is trustworthy to contract C, and accepts vulnerability to M’s actions, then H trusts M contractually to C. The objective of H in trusting M is to anticipate that M will maintain C in the presence of uncertainty, and consequently, trust does not exist if H does not perceive risk.

Previously, we note that a user’s “anticipation”, defined as the user’s belief that the AI will work ‘as intended’, is a key aspect of interpersonal trust that we will use to define Human-AI trust. Indeed, “anticipation” happens when a user believes that the AI is capable of maintaining the contract, i.e., that the AI is trustworthy to the contract, which is a prerequisite of Human-AI trust.

**Warranted and unwarranted Human-AI trust.** H’s trust in M (to C) is warranted if it is caused by trustworthiness in M. This holds if it is theoretically possible to manipulate M’s capability to maintain C, such that H’s trust in M will change. Otherwise, H’s trust in M is unwarranted.

**Human-AI distrust.** If H (human) perceives that M (AI) is not trustworthy to contract C, and therefore does not accept vulnerability to M’s actions, then H distrusts M contractually to C. We say that it is warranted distrust if the distrust is caused by the non-trustworthiness of M.

---

**6 CAUSES OF TRUST**

The next natural question to ask is on the cause behind trust in an AI model. As established earlier, we say that trustworthiness is a prerequisite to warranted trust. What causes a model to be trustworthy? And what enables trustworthy models to incur trust? We divide causes of warranted trust into two types: intrinsic and extrinsic.\(^9\)

---

**6.1 Intrinsic Trust**

A model is more trustworthy when the observable decision process of the model matches user priors on what this process should be. This is equivalent to, for example, a doctor which is considered more trustworthy because they are citing various respectable studies to justify their claims.

Explanation in AI aims to explain the decision process of the AI to the user, such that they can understand why and how the decision was made. However, the process of explaining does not, in itself, enable intrinsic trust. Only when (1) the user successfully comprehends the true reasoning process of the model, and (2) the reasoning process of the model matches the user’s priors of agreeable reasoning, intrinsic trust is gained.

---

\(^7\)Not all researchers and organizations may be interested in the distinction between warranted and unwarranted trust, unless required by regulation. We contend that this is an ethical issue. In this work, we assume that it is a core motivation of the field to remain ethical and not implement or prioritize unwarranted trust.

\(^8\)We consider AI as an automation attributed with intent, and thus, Human-AI trust is a sub-case of human-machine trust [28, 55]. Trust in automation which is not anthropomorphized can be considered as reliance (§9.1).

\(^9\)The distinction, and choice of terminology, are ours. Note that this is unrelated to ‘intrinsic or enforceable’ trust as coined by Hofstede [29].
Alon Jacovi, Ana Marasović, Tim Miller, Yoav Goldberg.

USER OBSERVES (TRUSTWORTHY EVALUATION OF) MODEL BEHAVIOR ARE TRUSTWORTHY

REASONING PROCESS ALIGNS WITH HUMAN REASONING

USER OBSERVES (INTERPRETATION/EXPLANATION OF) REASONING PROCESS

EXTERNAL SYMPTOMS OF MODEL BEHAVIOR ARE TRUSTWORTHY

USER GAINS INTRINSIC TRUST

USER GAINS EXTRINSIC TRUST

Figure 2: A schematic of the causes of warranted trust in AI models (summary of §6).

For example, a decision tree is a model whose inner workings can be well-understood by the user (if it is sufficiently small). However, e.g., for a task involving complex expert knowledge, a layman user will not be able to gain intrinsic trust in the model regardless of how ‘simple’ and interpretable the model is. If the user has no prior on what behavior is trustworthy for the given task, intrinsic trust will not be gained, even if the AI is easy to understand.

Further of note is the granularity of interpretability that will facilitate intrinsic trust. The issue of enumerating relevant priors to the user through explanation is not trivial: it is possible to convey information about the reasoning process which is accurate to the model and easy to comprehend, but nevertheless irrelevant to the priors that the user cares to know about. In the same way, it may be possible to derive an explanation of the reasoning process at very coarse granularity and remain descriptive of the relevant priors.

In the example of the doctor citing respectable studies, it is not necessary for the user to be able to understand those studies—merely that they are respectable, and that their claims match the claims of the doctor. As another example, “the model should not make any decision based on the number of commas in the text” can be such a prior, if the user truly perceives “the number of commas” as irrelevant to the decision. Generally, we regard any prior as relevant so long as it is useful to upholding the contract in question. For example, the absence of heuristical behavior (such as detecting the number of commas) can be such a prior towards a contract of model correctness.

Intrinsic trust can be increased in a disciplined manner by formalizing the priors behind trustworthy or non-suspicious behavior, and incorporating behavior which upholds those priors. The documents outlined in Table 1 (§3) could be useful for that.

6.2 Extrinsic Trust

It is additionally possible for a model to become trustworthy not through explanation, but through behavior: in this case, the source of trust is not the decision process of the model, but the evaluation methodology or the evaluation data. This is equivalent to a doctor who is considered more trustworthy because they have a long history of making correct diagnoses; or because they graduated from a prestigious institute that is considered to have rigorous student evaluation. The trust comes from observing symptoms of a trustworthy model, rather than the model’s inner working.

Extrinsic trust can be seen as the inverse of intrinsic trust: where aligned priors cause a trustworthy model, a trustworthy model causes convincing external symptoms. Since both share a causal relationship with a trustworthy model, they are both utilized by people as indicators of trustworthiness and thus incur warranted trust. This perspective is outlined in Figure 2.

For AI models, extrinsic trust is, in essence, trust in the evaluation scheme. To increase extrinsic trust is a matter of justifying that a model can generalize to unseen instances based on expected behavior of the model on other unseen instances.

Extrinsic trust is gained by two independent requirements: (1) when the model is trustworthy, and (2) when the evaluation scheme is trustworthy. To show that some evaluation scheme is ‘trustworthy’ is to justify that the distribution of instances during evaluation matches the distribution of the true unseen instances that will require trust (in a specific contract) by the user in the future—or in other words, guarantee that it is only possible for the AI to pass the evaluation if it is capable of maintaining the contract.

Three main methods of evaluation towards extrinsic trust:

1. **By proxy.** Expert (human) opinion on AI reasoning or behavior can enable non-experts to gain extrinsic trust in the AI. Note that the expert does not necessarily gain trust at this point, because the expert may or may not be vulnerable to the AI’s decisions, as the interaction between the expert and the AI is made under different conditions.

2. **By observing performance data.** An evaluation scheme that verifies whether the AI does not discriminate against a sub-population may be different from an evaluation scheme that verifies general performance ability.

3. **By formalizing the evaluation scheme.** A formal evaluation scheme can be used to ensure that the AI is capable of maintaining the contract in the future.
terms than that between the AI and the user. Also of a note is that what exactly constitutes a trustworthy expert for this purpose is a question of interpersonal trust, and not HUMAN-AI trust.

(2) Post-deployment data. Most simply, the examples that the model sees during production are the most trustworthy representatives of general behavior evaluation, notwithstanding issues of distribution shift over time. Such examples may or may not have gold supervision, and performance may be measured by some weaker signal. Although the distribution of these examples is realistic, it is not controllable, and thus there is little control on the specification of contracts to be evaluated.

(3) Test sets. Sets of examples, distributed in some specific way, for which gold labels are available. Test sets are generally seen as imperfect approximators of post-deployment data, as their distribution is not representative of true unseen data [12] and thus cannot imply overall generalization.

However, this perspective only takes into account one type of contract: high general post-deployment performance. The distribution of the test examples is controllable, such that they can specialize in specific contracts, such as specific phenomena, robustness and adversarial robustness, fairness and ethics, privacy, and so on. For example, by showing that a model performs equally for two large (and convincingly distributed) collections of people of different races, a user may deem the model more trustworthy to the contract: “I trust you to not discriminate based on race.” Previous work has used specialized test sets to verify particular contracts [5, 36, 69], or outlined methodologies for constructing such evaluation [17, 33, 58, 63].

How can we verify whether an evaluation scheme (in particular, test sets and deployment data) is trustworthy? Using data for validation assumes the following: (1) that the data accurately represents the underlying mechanism it comes from (e.g., the label is correct, the distribution is representative for the contract at the time of data collection); (2) that the underlying mechanism is negligibly affected by distribution shift over time; and (3) that the evaluation metrics represent the contract—i.e., that a positive result implies the capability to maintain the contract, and the inverse. The degree to which these assumptions hold affects the validity of the evaluation scheme.

Notably, point (3) is affected by the ‘accurate’ formalism of contracts. For example, there are multiple formal measures of fairness such as individual fairness, demographic parity and equalized error rates [73]. However we cannot say that each one of these measures completely encapsulates the social requirement of fairness: each measure formalizes a different aspect of fairness, and there cannot be a solution that satisfies all of them.

7 EXPLAINABILITY AND TRUST

As mentioned, the following is a common claim of XAI literature:

XAI for Trust (common): A key motivation of XAI and interpretability is to increase the trust of users in the AI.

However, the above claim is difficult to probe without a definition of trust. Following the formalization in this work, we can now elucidate this claim with details on the true nature of this motivation:

XAI for Trust (extended): A key motivation of XAI and interpretability is to (1) increase the trustworthiness of the AI, (2) increase the trust of the user in a trustworthy AI, or (3) increase the distrust of the user in a non-trustworthy AI, all corresponding to a stated contract, so that the user develops warranted trust or distrust in that contract.

Let us clarify this claim by unraveling it:

A key motivation of XAI and interpretability is...

... to (1) increase the trustworthiness of the AI...

AI is said to be trustworthy to a contract if it is capable of maintaining the contract. Then XAI is a method of creating a capability by revealing the relevant signals in the AI reasoning process (as in the example of the random baseline-like model, §3.1). An AI model that hides these signals would be less trustworthy as they fail to uphold some contracts.

... (2) increase the trust of the user in a trustworthy AI...

The goal of developing trust, from the user’s perspective, is to enable the ability to anticipate behavior in the presence of risk. Then XAI is a method of allowing the user easier access to the signals that enable this anticipation.

... or (3) increase the distrust of the user in a non-trustworthy AI...

Similarly, the user’s goal in distrust, in the presence of risk, is to anticipate when desired behavior will not happen. This is the inverse of (2), therefore XAI is a method of enabling distrust.

... all corresponding to a stated contract...

The above is only relevant with respect to specific contracts that dictate what precisely is anticipated. Therefore, the contract must be explicitly recognized by the XAI methodology, so that it reveals information which is relevant to create or reveal the capability of the AI to maintain the contract.

... so that the user develops warranted trust or distrust in that contract.

For the user to achieve their goal of anticipation, their trust should be warranted, and XAI verifies this by revealing the intrinsic or extrinsic sources (causes) of the trust.

8 EVALUATING TRUST

The last question that remains for us to discuss is on the evaluation of trust. What should evaluation of trust satisfy? Do current methods of evaluating trust fulfill these requirements?

8.1 Vulnerability in Trust and Trustworthiness

The distinction of evaluation between trust and trustworthiness stems from their distinction on vulnerability: trustworthiness does not require the trustor to accept risk to manifest, and trust does. For instance, the evaluation of the accuracy of AI models may be used to increase external trust in the model for others by asserting trustworthiness or by providing evidence for trustworthiness, but they are not on their own evaluating any notion of trust, as there is no risk involved.
Vulnerability in Trust. By definition, the question of trust does not exist when the user does not assume a risk in trusting the AI. The user must depend on the AI on some level for risk to manifest: for example, when the user is already confident in the answer and does not actually depend on the AI—as is the case for many machine learning datasets today—there is no unfavorable event that directly results from the AI’s actions. Therefore, experiments which simply ask the users whether they trust the model for some trivial task evaluate neither trust nor trustworthiness.

We illustrate this point with the ERASER benchmark [13] as a case-study (Figure 3). ERASER is a benchmark for natural language processing interpretability, consists of multiple datasets. For the purpose of this illustration, we set very simple measures of vulnerability and required effort for humans to solve associated tasks. Foremost, we assume the task domain (what the AI is trained on) is also the use-case (the real interaction setup), e.g., the user must provide the label for a problem instance, and is advised by an AI model trained to do the same. We heuristically measure required human effort using the length of the input (average number of tokens). The longer the text is, the more time-consuming and laborious is the task to people, and consequently people are more dependent on AI. We set the “effort boundary” around 774 tokens—a few of minutes of reading—and the “vulnerability boundary” right from natural language inference [7], commonsense QA [70], and multi-sentence reading comprehension [34], as these tasks are designed only to test models’ capabilities in capturing particular linguistic phenomena, and, to the best of our knowledge, there is no undesirable event from interacting with them in the base scenario.

The question of trust can be considered meaningful for tasks above the “effort boundary” and right from the “vulnerability boundary”, and notably, only two to three ERASER tasks can be placed in those areas (e.g., evidence inference).11

We advocate that only use-cases that can be attributed with both considerable required human effort and vulnerability, are used to target, evaluate and discuss trust. Although in our ERASER example we treat the task domains as use-cases, comprehensive discussion in this area must develop the use-case explicitly. The use-case may not necessarily be an exact replica of the AI’s task domain, and thus, the question of vulnerability would depend entirely on the use-case; additionally, we note that the measure of difficulty is not solely constrained to how time-consuming the task is, as even ‘easy’ tasks may require trust if the stakes are high.

Vulnerability in Trustworthiness. Whether a model is intrinsically or extrinsically trustworthy is unrelated to the existence of vulnerability in the user. For example, a domain expert can verify the intrinsic trustworthiness of a model by verifying whether the model performs the expected reasoning steps to arrive at its decision, despite the expert not necessarily assuming any vulnerability in doing this inspection, the same way that a class teacher evaluates their students.

---

11 Given a scientific article describing a randomized controlled trials, evidence inference is the task of inferring whether a given intervention is reported to either significantly increase, significantly decrease, or have no significant effect on a specified outcome, as compared to a comparator of interest.
8.2 Warranted and Unwarranted Trust

Of course, it is impossible to differentiate between warranted and unwarranted trust simply by evaluating whether the user trusts the model. In this area, Kaur et al. [32] show a synthetic experimental setup to evaluate unwarranted trust, and conclude that even data scientists are susceptible to develop unwarranted trust, despite some mathematical understanding of the models. Smith-Renner et al. [62] show similar conclusions on unwarranted trust in a different experimental setup. However, this area is underdeveloped.

Our discussion of warranted trust in Sections 4 and 5 suggests a possible evaluation based on manipulationist causality—i.e., that if the trust is warranted, the level of trust can be manipulated by manipulating the cause of trustworthiness. This gives rise to the following evaluation protocol:

1. Measure the level of trust in an interaction.
2. Manipulate the real trustworthiness of the model (e.g., by handicapping it in some way; by improving its predictions; or even by replacing the model with an oracle).
3. Measure the level of trust after the manipulation.

The amount of change due to the manipulation indicates the level of warranted trust.

8.3 Evaluating ‘Anticipation’

While not an evaluation of trust, there are methods of evaluating the ability of users to successfully anticipate the AI’s behavior, an important aspect of Human-AI trust, via simulatability [15, 25]—the ability of the user to simulate the outcome of the AI on an instance level. Thus, simulatability can serve as one proxy signal to assess whether the goal of trust has been achieved, though we note that it does not concretely verify the existence of trust—on the contrary, it relies on a prior assumption that trust exists. Additionally, since simulatability is performed on an instance level, it does not clarify contracts that are not behaviorally observable on an instance level (e.g., contracts that deal with privacy concerns, or with global properties of the predictions over a large sample set).

9 DISCUSSION

The basis of our formalization of Human-AI trust is the basic definition of interpersonal trust in sociology (§2). In this section, we first discuss additional aspects of interpersonal trust and human-machine trust (trust in automation), and their relevance to our definitions (§5). We then present two directions for extension of our formalization inspired by other factors of interpersonal trust.

9.1 Other Elements of Interpersonal Trust and Trust in Automation

Trust vs Reliance. Baier [4] introduces the term reliance for trust in inanimate objects, such as trusting a stool to hold your weight. We may feel betrayed when our trust fails us, but this is not true for reliance because we did not attribute intent to the trustee [68]. Despite the fact that AI models are tools, people anthropomorphize them and assign intent to them [30, 52]. This makes us believe that reliance is not suitable for AI models, and for this reason we aim to define Human-AI trust as a distinct sub-case of human-machine trust, or trust in automation, where otherwise the automation may not be attributed with intent (in which case, it is reliance).

Warranted and Justified Trust in Social Science. Our definition of warranted Human-AI trust (§5) is trust which is caused by trustworthiness (to some contract). This definition is only applicable to Human-AI trust, and is not strictly relevant in sociology.

Sociology elects to define trustworthiness and justified trust by the effect, rather than the cause: in interpersonal trust, the trustee is trustworthy, and the trust in them was justified, if the trust was not betrayed. Two natural questions emerge: (1) why does sociology define interpersonal trust in this way? (2) and why do we not adopt this definition for Human-AI trust?

The capability of the trustee to maintain the trust, and their intent to do so, are both prerequisites to interpersonal trust, but are not necessarily sufficient. This is due to the complex nature of human interaction, with respect to elements of chance and outside factors (e.g., the difference between innocent mistake and negligence, ‘chains’ of trust, and so on). This makes the assignment of blame difficult or ill-defined: e.g., if the trustee was fully capable of maintaining their trust, and intended to do so, but contracted a common cold which ultimately prevented them from achieving their goal—it is difficult to define them as trustworthy or otherwise.

Human-AI trust does not share these limitations, as (1) the AI (by definition as an automation) does not possess real intent; and (2) the AI’s capabilities are well-defined. As a result, we diverge from sociology definitions on trustworthiness and justified trust, and adopt a stricter causal definition of warranted trust.

Can Trust Become Warranted? In this work, we discuss warranted trust as a ‘state’ that trust can become, on a binary. This is equivalent to the notion of trust becoming ‘calibrated’ as discussed by Lee and See [43]. However, this is not realistic since trust can increase or decrease freely—on a sliding scale. But is the sliding scale, in itself, sufficient to describe trust? More specifically, Hoffman [28] argues that trust cannot be positioned on a binary, or even a sliding scale, as trust is multi-dimensional. In his own words:

In my own relation to my word processing software, I am positive that it will perform well in the crafting of simple documents, but I am simultaneously confident it will crash when the document gets long, or when it has multiple high-resolution images. And every time that there is a software upgrade, the trusting of many of the functions becomes tentative and skeptical. [Therefore,] trust is not a single state.

We argue that Hoffman’s view is implicitly informed by a notion of contractual trust. In other words, while general trust cannot be described by a single state (or sliding scale) of warranted or unwarranted, trust in a contract can, since it is ‘calibrated’ by the capability of the model to maintain the contract, where contracts are dimensions.

9.2 Future Extensions of the Formalization

Trust in the AI Developer. Literature in sociology often considers aspects of larger social constructs (beyond a single transaction of two people), specifically relationships and communities and the propagation of trust in them. A common theme in such models
We discuss intrinsic reasoning and extrinsic behavior as causes of useful and valuable as methods of assessing this property, it is dangerous to rely on them solely.  
(4) Trust is only ethically desirable if it is warranted. Unwarranted trust is not guaranteed to accomplish its goal, since it is not sourced by trustworthiness. This can cause issues of abuse, disuse or misuse of the AI. While unwarranted trust may be in the interest of some parties, AI research should make efforts to both diagnose and avoid unwarranted trust by, among other things, identifying relevant contracts and assessing trustworthiness.  
(5) Distrust is not strictly undesirable if it is warranted. Just as trustworthiness and warranted trust must both manifest for the AI to be useful in practice, so too must warranted distrust follow non-trustworthiness for contracts which are relevant to the application. Completely trustworthy AI, for all relevant contracts, may be prohibitively difficult to achieve—in which case warranted distrust is the mechanism that will allow the imperfect AI to be useful.  
(6) Explanation seems to be uniquely positioned for Human-AI trust as a method for causing intrinsic trust for general users. Other causes of trust, such as empirical evaluation and authority, are extrinsic. This may help to explain the recent interest in XAI.  
(7) Methods of evaluating whether trust is warranted are underdeveloped, and require future work. It is important to verify whether the trust of the user in an AI is warranted. However, there is currently little work on achieving this goal, and this question is positioned to be central for future research on Human-AI trust.  

We hope that this work informs AI research in the following ways: (1) by encouraging more accurate discussion of trust in AI, through a more transparent definition of trust which includes the notions of contractual and warranted trust; (2) by encouraging claims on trust and on methods of causing or evaluating trust to be founded on, and distinguish between, the notions of intrinsic and extrinsic trust; (3) by requiring to explicitly recognize and verify risk in user studies that make claims on trust.
