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 promote them, 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, interpersonal trust (i.e., trust between people) as defined by sociologists. This model rests on two key properties: 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 model is trust that some implicit or explicit contract will hold, and a formalization of 'trustworthiness' (that detaches from the notion of trustworthiness in sociology), and with it concepts of 'warranted' and 'unwarranted' trust. We 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

trust, distrust, trustworthy, warranted trust, contractual trust, artificial intelligence, sociology, formalization

FAccT '21, March 3–10, 2021, Virtual Event, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8309-7/21/03…$15.00
https://doi.org/10.1145/3442188.3445923

ACM Reference Format:
Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442188.3445923

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 AI [8, 11, 76, 81]. The recent rapid growth in *explainable AI* (XAI) is also, in part, motivated by the need to maintain trust between the human user and AI [32, 36, 47, 55, 62, 77]. By designing AI that users can and will trust to interact with, 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 human trust in AI? For what goals does the cognitive mechanism of trust exist? How can we design AI that 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 that can be integrated in society. We consider 'artificial intelligence' to be any automation that is attributed with intent by the user [social attribution, 55], 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 aspects of trust that 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 model, based on the *prerequisites* and *goals* of the trustor gaining trust in the AI (Section 2). Specifically, the trustor must be *vulnerable* to the agent'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 model being trusted with (i.e., what is 'intended')?*; and (2) *what differentiates trust that achieves this goal, and trust that 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, the 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 using 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 of the ability to anticipate. Finally, in Section 9 we expand on other aspects of interpersonal trust and human-machine trust (automation not attributed with intent), their relation to our notion of Human-AI trust, and possible future extensions of our formalization.

Contributions. We provide a formal perspective of Human-AI trust that is rooted in, but nevertheless not the same as, interpersonal trust as defined by sociologists. 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 nuance behind our choice of formalization, made available in §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 [52]. The goal of trust is to "make social life predictable [by anticipating the impact of behavior], and make it easier to collaborate between people" [56].[1]

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 trustor's ability to anticipate the trustee—where 'anticipating' refers to a belief that the trustee will act in the trustor's best interests. We maintain that Human-AI trust exists for the same purpose, as a sub-case of trust in automation, following Hoffman [29]: 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 [25]. Therefore, "to act in A's best interest" is to avoid any unfavorable events. 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, AI-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 (and not certainly) be incorrect for trust to manifest. Similarly, from the side of the applicants (if they have a choice as to whether to use the AI model), 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 model 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 [71]. Importantly, distrust is *not* equivalent to the absence of trust [53], 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'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., 23, 47, 64]. What does this mean, exactly?

[1] This definition of trust is considered overly simplistic by many in sociology. In Section 9 we discuss aspects of more elaborate formalizations of interpersonal trust, and whether they are relevant to Human-AI trust.

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).[2] 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.[3] As an illustrative example, consider a credit-scoring AI model that 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 trustor 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 [27] and Tallant [71] refer to as *trust with commitment* or *contractual trust*. Contractual trust is when a trustor has a belief that the trustee will stick to a specific contract.[4]

In this work, we contend that all Human-AI trust is contractual,[5] and that regardless of what the contract is in a particular interaction, to discuss Human-AI trust, **the contract must be explicit**.

Generally, the contract may refer to any functionality that 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 [43, 47].

Contracts and contexts. The idea of context is important in trust: people can trust something in one context but not another [29]. 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).[6] 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 [57], reproducibility checklists [59], fairness checklists [50], 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 universal design.

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 or expected agreement, even in the case of a social contract.

[2] 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.

[3] E.g., calibrated probabilities [46], where the classification probabilities of a model are calibrated with some measure of its uncertainty, can produce this effect.

[4] To our knowledge Hawley [27] is the first to formalize trust as "trust with commitment [= contract]." Tallant [71] expands on their work with terminology of contractual trust.

[5] Although they do not refer to contractual trust in their work, Hoffman [29] 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.

[6] 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		**Documentations**	**Explanatory Methods/Analyses**
Key Requirements	*Factors*		
Human agency and oversight	· Foster fundamental human rights	Fairness checklists	See "Diversity, non-discrimination, fairness"
	· Support users' agency	All	User-centered explanations [62]
	· Enable human oversight	N/A	Explanations in recommender systems [42]
Technical robustness and safety	· Resilience to attack and security	Factsheets (security)	Adversarial attacks and defenses [21]
	· Fallback plan and general safety	N/A	N/A
	· A high level of accuracy	Model cards (metrics)	N/A
	· Reliability	Factsheets (concept drift)	Contrast sets [17], behavioral testing [61]
	· Reproducibility	Reproducibility checklists	"Show your work" [14]
Privacy and data governance	· Ensure privacy and data protection	Datasheets/statements	Removal of protected attributes [60]
	· Ensure quality and integrity of data	Datasheets/statements	Detecting data artifacts [24]
	· Establish data access protocols	Datasheets/statements	N/A
Transparency	· High-standard documentation	All	N/A
	· Technical explainability	Factsheets (explainability)	Saliency maps [65], self-attention patterns [41], influence functions [39], probing [16]
	· Adaptable user-centered explainability	Factsheets (explainability)	Counterfactual [22], contrastive [54], free-text [28, 51], by-example [39], concept-level [20] explanations
	· Make AI systems identifiable as non-human	N/A	N/A
Diversity, non-discrimination, fairness	· Avoid unfair bias	Fairness checklists	Debiasing using data manipulation [70]
	· Encourage accessibility and universal design	N/A	N/A
	· Solicit regular feedback from stakeholders	Fairness checklists	N/A
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 [69]
	· Assess the impact on society and democracy	Fairness checklists	Explanations designed for applications such as fact checking [3] or fake news detection [48]
Accountability	· Auditability of algorithms/data/design	Factsheets (lineage)	N/A
	· Minimize and report negative impacts	Fairness checklists	N/A
	· Acknowledge and evaluate trade-offs	N/A	Reporting the robustness-accuracy trade-off [1] or the simplicity-equity trade-off [38]
	· Ensure redress	Fairness checklists	N/A

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) [68, 75, 80], 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 that 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* [53]. Warranted trust is sometimes referred to as trust that is *calibrated* with trustworthiness [44]. 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 model 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 system development. 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 (Figure 1). For instance, Ghassemi et al. [19] show a case where the interface can increase doctors' confidence in a tool, despite not significantly increasing their accuracy.

Formally, we define warranted Human-AI trust via a causal (interventionist) relationship with trustworthiness: incurred Human-AI

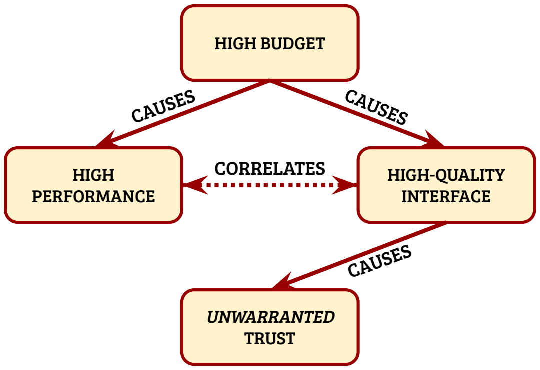

Figure 1: An example of causes of trust, in the context of warranted and unwarranted trust. For a contract of model performance, *high-quality interface* is not a cause of high performance, and therefore, any trust that the user gains as a result of the interface is unwarranted.

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 that is warranted.[7] Therefore, we contend that **when pursuing Human-AI trust, unwarranted trust should be explicitly evaluated against, and avoided or otherwise minimized.** See [58] 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 an AI model is incapable of maintaining some relevant contract, it is a desired outcome (desired by the developer) that the user develop warranted distrust in that contract**, that will be beneficial in applying the AI model to some scenario despite 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 that is attributed with human-like intelligence by the human interacting with it.[8]

Trustworthy AI. An AI model 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" occurs when a user believes that an AI model is capable of maintaining the contract, i.e., a user believes that the model 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 that 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.

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.**

[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 [29, 58]. Trust in automation that 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 [30].

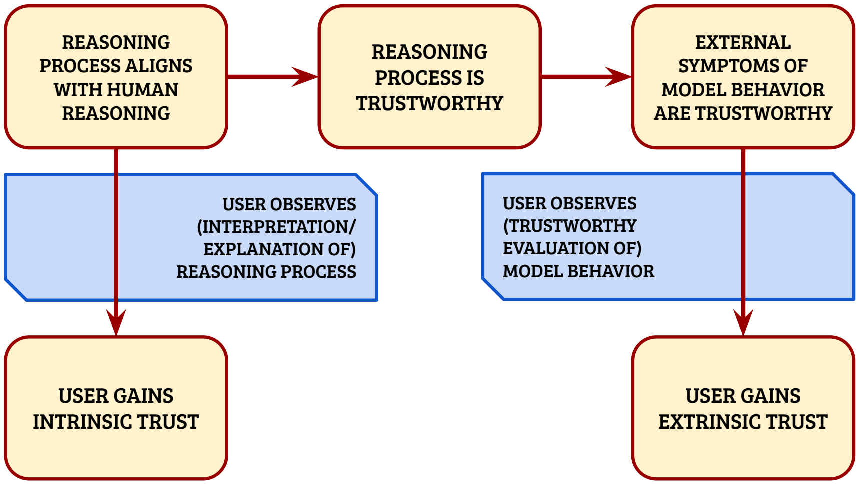
REASONING PROCESS ALIGNS WITH HUMAN REASONING
REASONING PROCESS IS TRUSTWORTHY
EXTERNAL SYMPTOMS OF MODEL BEHAVIOR ARE TRUSTWORTHY
USER OBSERVES (INTERPRETATION/ EXPLANATION OF) REASONING PROCESS
USER OBSERVES (TRUSTWORTHY EVALUATION OF) MODEL BEHAVIOR
USER GAINS INTRINSIC TRUST
USER GAINS EXTRINSIC TRUST

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

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 that 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 that upholds those priors. The documentations 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 used 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.[10]

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 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

[10] E.g., 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.

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, 37, 73], or outlined methodologies for constructing such evaluation [17, 34, 61, 67].

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 [78]. 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 (e.g., Table 1).

... (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 that 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. 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 that simply ask the users whether they trust the model for some trivial task evaluate *neither* trust nor trustworthiness.**

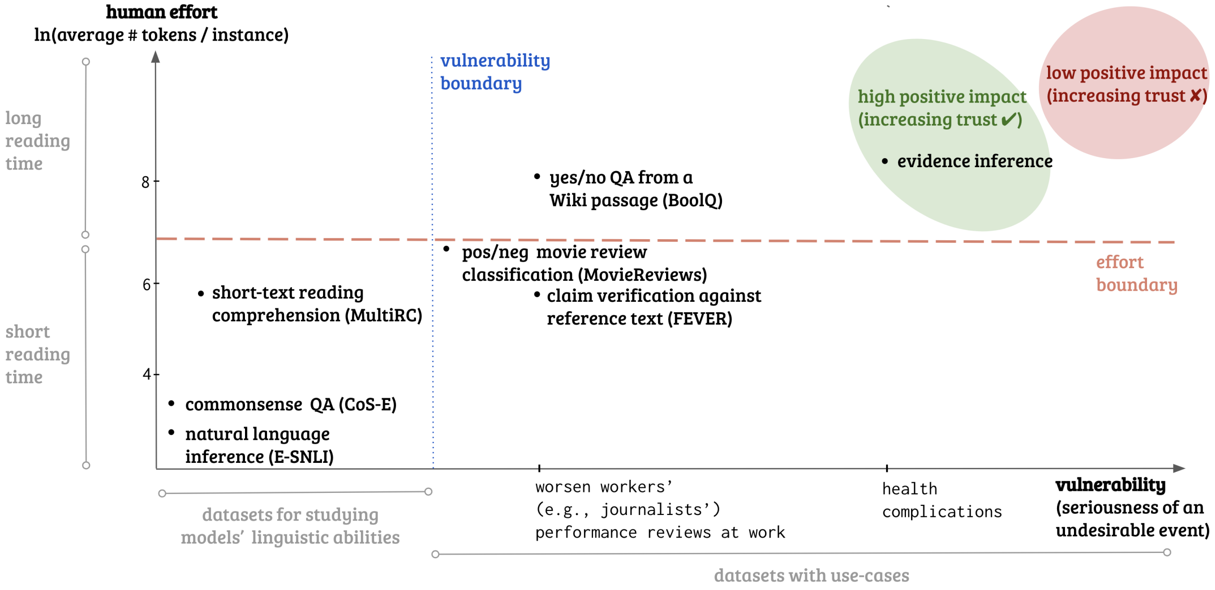

Figure 3: Categorization of datasets that are commonly used to advance interpretable NLP with respect to required effort from people to complete associated tasks (y-axis) and seriousness of an undesirable event (x-axis). Datasets that require long reading time and that have associated use-cases are better suited for studying trust. The average number of tokens per instance are: e-SNLI (16), CoS-E (28), MultiRC (303), FEVER (327), MovieReviews (774), BoolQ (3583), Evidence Inference (4761).

We illustrate this point with the ERASER benchmark [13] as a case-study (Figure 3). ERASER is a benchmark for natural language processing interpretability, and consists of multiple datasets. For the purpose of this illustration, we set simple measures of vulnerability and required effort for humans to solve associated tasks. We also assume that 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), expecting that the longer the text is, the more time-consuming and laborious is the task to people, and consequently people are more dependent on AI [49]. We set the "effort boundary"—a threshold for tasks requiring little effort—around 774 tokens (a few minutes of reading). Motivated by categorization in [63], we set the "vulnerability boundary"—a hypothetical threshold for tasks without an associated undesirable event— right from natural language inference [7], commonsense QA [74], and multi-sentence reading comprehension [35], 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.

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. [33] 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. [66] show similar conclusions on unwarranted trust in a different experimental setup. However, this area is underdeveloped.

[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.

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, 26]—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 from 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 [72]. Despite the fact that AI models are tools, people anthropomorphize them and assign intent to them [31, 55]. This makes us believe that reliance is not suitable for AI models, and for this reason we aim to define Human-AI *trust* instead of reliance. This positions 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 that 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 caught a common cold that 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 [44]. 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 [29] 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 is the incentives of both parties (e.g., trust in family vs trust in a business partner).

This notion is, of course, relevant to Human-AI trust. For example, since there is a close relationship between workplaces with discriminatory practices and discriminatory AI tools [79], it is likely that those who are discriminated against have more incentive to trust AI tools produced by teams that represent them. By recognizing the incentives of the developer, the user may gain trust "in the AI", separately from the other causes described in this work.

AI as an automation does not embody intent. Formally, the intent of the AI developer manifests in the capability of the model to maintain specific contracts, rather than adopting some anthropomorphized notion of intent. Therefore, we make two claims on the nature of this trust.

Trust in the AI model based on trust in the AI developer is an instance of *interpersonal* trust by proxy, and not Human-AI trust. Therefore, studies of the influence of the trustee's incentives on Human-AI trust, should build on the existing research in sociology that investigates the influence of relationships and communities on interpersonal trust [45]. Consequently, **the question of whether this trust is *warranted* or *unwarranted* is ill-defined.** Our definition of warranted trust—as trust that is *caused* by trustworthiness—is a definition that does not apply to interpersonal trust (as mentioned in §9.1).

To conclude, while trust in the AI developer *could* possibly influence the user-AI interaction, and should therefore be studied and modeled, it is strictly not part of our model of Human-AI trust. We therefore position this topic as future work towards a more nuanced model of trust in the interaction between people and AI.

Personal Attributes of the Trustor. As mentioned, in this work we consider each interaction as a 'clean slate' transaction of trust. Future work in this area may incorporate elements of the personal attributes of the trustor into the model, such as personality, socio-cultural background [40], prior existence of trust or distrust, or the restoration of trust that has been betrayed.

10 CONCLUSION

While Trust is a central component in the interaction between a user and the AI, current references to trust in the XAI literature are vague and under-specified. In this work, we provide several working definitions for trust-related terms, and answer the questions of what is necessary to allow trust to occur, and what is necessary for the goal of trust (anticipation of desired contracts) to be achieved. The goal of this formalization is to allow a disciplined method of developing trustworthy AI models that will incur trust in practice. We discuss intrinsic reasoning and extrinsic behavior as causes of warranted trust to be pursued in designing a trustworthy model. This is directly connected to XAI, which can provide the framework to verify whether a model is trustworthy or to create trustworthiness in the model. We further note that the question of trust in this context hinges on a notion of vulnerability, which is not satisfied by many evaluation methods currently used in XAI.

Takeaways. We collect the various conclusions in this work into guidelines that should inform the design of AI that is both trustworthy and trusted.

1. **The assessment of risk is necessary prior to the assessment of trust.** When deciding whether an AI requires trust, or when evaluating trust, verify the existence of vulnerability of the user in the actions of the AI, by: (1) verifying that the user considers some of the actions of the AI as unfavorable to them; and (2) verifying that the user considers both the favorable and unfavorable outcomes as realistic.
2. **AI developers should be explicit about the contracts that their models maintain.** AI developers should use the right affordances to make the relevant contracts explicit. This can help to avoid situations where a contract is implicitly assumed by a user, despite the developer not considering the contract to be upheld (unwarranted trust).
3. **Successful anticipation, while the goal of trust, is *not* indicative of warranted trust.** The trust can be unwarranted if it is not sourced in trustworthiness, in which case, the anticipation may depend on a different variable that does not exist in other situations (such as the quality of the AI user interface). Therefore, though simulatability methods are 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, we assert that 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 that 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 questioned 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 that 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.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful feedback. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant agreement No. 802774 (iEXTRACT) and from the Australian Research Council (ARC) Discovery Grant DP190103414: *Explanation in Artificial Intelligence: A Human-Centred Approach.*

REFERENCES

[1] David Alvarez Melis and Tommi Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). 7775–7784. http://papers.nips.cc/paper/8003-towards-robust-interpretability-with-self-explaining-neural-networks.pdf

[2] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, Darrell Reimer, John T. Richards, Jason Tsay, and Kush R. Varshney. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM J. Res. Dev.* 63, 4/5 (2019), 6:1–6:13. https://doi.org/10.1147/JRD.2019.2942288

[3] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

[4] Annette Baier. 1986. Trust and antitrust. *ethics* 96, 2 (1986), 231–260.

[5] Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Sentiment Analysis Is Not Solved! Assessing and Probing Sentiment Classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy, 12–23. https://doi.org/10.18653/v1/W19-4802

[6] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Trans. Assoc. Comput. Linguistics* 6 (2018), 587–604. https://transacl.org/ojs/index.php/tacl/article/view/1464

[7] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 632–642. https://doi.org/10.18653/v1/D15-1075

[8] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[9] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An Interpretable Model with Globally Consistent Explanations for Credit Risk. *CoRR* abs/1811.12615 (2018). arXiv:1811.12615 http://arxiv.org/abs/1811.12615

[10] F. Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, A. Bau, and James R. Glass. 2019. What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. In *AAAI*.

[11] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. arXiv:2006.11371 [cs.CV]

[12] Harm de Vries, Dzmitry Bahdanau, and Christopher D. Manning. 2020. Towards Ecologically Valid Research on Language User Interfaces. *CoRR* abs/2007.14435 (2020). arXiv:2007.14435 https://arxiv.org/abs/2007.14435

[13] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408

[14] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2185–2194. https://doi.org/10.18653/v1/D19-1224

[15] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[16] Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics, Berlin, Germany, 134–139. https://doi.org/10.18653/v1/W16-2524

[17] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating NLP Models via Contrast Sets. *CoRR* abs/2004.02709 (2020). arXiv:2004.02709 https://arxiv.org/abs/2004.02709

[18] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for Datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[19] Marzyeh Ghassemi, Mahima Pushkarna, James Wexler, Jesse Johnson, and Paul Varghese. 2018. ClinicalVis: Supporting Clinical Task-Focused Design Evaluation. *CoRR* abs/1810.05798 (2018). arXiv:1810.05798 http://arxiv.org/abs/1810.05798

[20] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 9273–9282. http://papers.nips.cc/paper/9126-towards-automatic-concept-based-explanations

[21] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6572

[22] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations *(Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 2376–2384. http://proceedings.mlr.press/v97/goyal19a.html

[23] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. https://doi.org/10.1145/3236009

[24] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 107–112. https://doi.org/10.18653/v1/N18-2017

[25] Sven Ove Hansson. 2018. Risk. In *The Stanford Encyclopedia of Philosophy* (fall 2018 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[26] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *CoRR* abs/2005.01831 (2020). arXiv:2005.01831 https://arxiv.org/abs/2005.01831

[27] Katherine Hawley. 2014. Trust, distrust and commitment. *Noûs* 48, 1 (2014), 1–20.

[28] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding Visual Explanations. In *ECCV*. https://arxiv.org/abs/1807.09685

[29] Robert R Hoffman. 2017. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive systems engineering: The future for a changing world* (2017), 137–163.

[30] Gert Jan Hofstede. 2006. Intrinsic and Enforceable Trust: A Research Agenda. *European Association of Agricultural Economists, 99th Seminar, February 8-10, 2006, Bonn, Germany* (01 2006).

[31] Alon Jacovi and Yoav Goldberg. 2020. Aligning Faithful Interpretations with their Social Attribution. *CoRR* abs/2006.01067 (2020). arXiv:2006.01067 https://arxiv.org/abs/2006.01067

[32] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

[33] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna M. Wallach, and Jennifer Wortman Vaughan. 2019. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning.

[34] Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[35] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 252–262. https://doi.org/10.18653/v1/N18-1023

[36] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). arXiv:1711.11279 [stat.ML]

[37] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, Malvina Nissim, Jonathan Berant, and Alessandro Lenci (Eds.). Association for Computational Linguistics, 43–53. https://doi.org/10.18653/v1/s18-2005

[38] J. Kleinberg and S. Mullainathan. 2019. Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability. *Proceedings of the 2019 ACM Conference on Economics and Computation* (2019).

[39] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions *(Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 1885–1894. http://proceedings.mlr.press/v70/koh17a.html

[40] Hana Kopecka and Jose M Such. 2020. Explainable AI for Cultural Minds. https://sites.google.com/view/dexahai-at-ecai2020/home Workshop on Dialogue, Explanation and Argumentation for Human-Agent Interaction , DEXAHAI ; Conference date: 07-09-2020.

[41] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4365–4374. https://doi.org/10.18653/v1/D19-1445

[42] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[43] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[44] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[45] J David Lewis and Andrew Weigert. 1985. Trust as a social reality. *Social forces* 63, 4 (1985), 967–985.

[46] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. 1977. Calibration of probabilities: The state of the art. In *Decision making and change in human affairs*. Springer, 275–324.

[47] Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM* 61, 10 (2018), 36–43. https://doi.org/10.1145/3233231

[48] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 505–514. https://doi.org/10.18653/v1/2020.acl-main.48

[49] Brian Lubars and Chenhao Tan. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 57–67.

[50] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and H. Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

[51] Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2810–2829. https://doi.org/10.18653/v1/2020.findings-emnlp.253

[52] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[53] Carolyn McLeod. 2015. Trust. In *The Stanford Encyclopedia of Philosophy* (fall 2015 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[54] Tim Miller. 2018. Contrastive Explanation: A Structural-Model Approach. *CoRR* abs/1811.03163 (2018). arXiv:1811.03163 http://arxiv.org/abs/1811.03163

[55] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38. https://arxiv.org/abs/1706.07269

[56] B. Misztal. 1996. *Trust in Modern Societies: The Search for the Bases of Social Order*. Wiley. https://books.google.co.il/books?id=q3R1QgAACAAJ

[57] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. ACM, 220–229. https://doi.org/10.1145/3287560.3287596

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

[59] Joelle Pineau. 2020. The Machine Learning Reproducibility Checklist. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf.

[60] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7237–7256. https://doi.org/10.18653/v1/2020.acl-main.647

[61] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442

[62] Mireia Ribera and Àgata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *IUI Workshops*.

[63] D. Schlangen. 2020. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. (2020). https://arxiv.org/abs/2007.04792 arXiv:2007.04792.

[64] Philipp Schmidt and Felix Bießmann. 2019. Quantifying Interpretability and Trust in Machine Learning Systems. *CoRR* abs/1901.08558 (2019). arXiv:1901.08558 http://arxiv.org/abs/1901.08558

[65] K. Simonyan, A. Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *2nd International Conference on Learning Representations ICLR, Workshop Track Proceedings*. https://arxiv.org/abs/1312.6034

[66] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[67] Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2020. We Need to Talk About Random Splits. *CoRR* abs/2005.00636 (2020). arXiv:2005.00636 https://arxiv.org/abs/2005.00636

[68] Robert C. Solomon. 1998. Creating Trust. *Business Ethics Quarterly* 8, 2 (1998), 205–232. https://doi.org/10.2307/3857326

[69] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1679–1684. https://doi.org/10.18653/v1/P19-1164

[70] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1630–1640. https://www.aclweb.org/anthology/P19-1159

[71] Jonathan Tallant. 2017. Commitment in Cases of Trust and Distrust. *Thought: A Journal of Philosophy* 6, 4 (2017), 261–267. https://doi.org/10.1002/tht3.259 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/tht3.259

[72] Jonathan Tallant and Donatella Donati. 2020. Trust: from the Philosophical to the Commercial. *Philosophy of Management* 19, 1 (2020), 3–19.

[73] Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. oLMpics - On what Language Model Pre-training Captures. *CoRR* abs/1912.13283 (2019). arXiv:1912.13283 http://arxiv.org/abs/1912.13283

[74] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4149–4158. https://doi.org/10.18653/v1/N19-1421

[75] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2020. Trustworthy artificial intelligence. *Electronic Markets* (10 2020). https://doi.org/10.1007/s12525-020-00441-4

[76] Erico Tjoa and Cuntai Guan. 2019. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *CoRR* abs/1907.07374 (2019). arXiv:1907.07374 http://arxiv.org/abs/1907.07374

[77] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 272–283. https://doi.org/10.1145/3351095.3372834

[78] Suresh Venkatasubramanian. 2019. Algorithmic Fairness: Measures, Methods and Representations. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (Amsterdam, Netherlands) *(PODS '19)*. Association for Computing Machinery, New York, NY, USA, 481. https://doi.org/10.1145/3294052.3322192

[79] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems: Gender, race and power in AI. (2019). https://ainowinstitute.org/discriminatingsystems.pdf

[80] Stephen Wright. 2010. Trust and Trustworthiness. *Philosophia* 38, 3 (2010), 615–627. https://doi.org/10.1007/s11406-009-9218-0

[81] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges*. 563–574. https://doi.org/10.1007/978-3-030-32236-6_51