The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency

Bran Knowles  
John T. Richards  
Frens Kroeger  
b.h.knowles1@lancaster.ac.uk  
ajtr@us.ibm.com  
freds.kroeger@coventry.ac.uk  
Lancaster University  
TJ Watson Research Center, IBM  
Coventry University  
Lancaster, UK  
Yorktown Heights, New York, USA  
Coventry, UK

ABSTRACT
Efforts to promote fairness, accountability, and transparency are assumed to be critical in fostering Trust in AI (TAI), but extant literature is frustratingly vague regarding this “trust”. The lack of exposition on trust itself suggests that trust is commonly understood, uncomplicated, or even uninteresting. But is it? Our analysis of TAI publications reveals numerous orientations which differ in terms of who is doing the trusting (agent), in what (object), on the basis of what (basis), in order to what (objective), and why (impact). We develop an ontology that encapsulates these key axes of difference to a) illuminate seeming inconsistencies across the literature and b) more effectively manage a dizzying number of TAI considerations. We then reflect this ontology through a corpus of publications exploring fairness, accountability, and transparency to examine the variety of ways that TAI is considered within and between these approaches to promoting trust.

CCS CONCEPTS
• Human-centered computing → HCI theory, concepts and models; • Social and professional topics → Computing profession.

KEYWORDS
Trust, artificial intelligence, fairness, accountability, transparency

1 INTRODUCTION
The matter of trust is strongly associated with considerations of fairness, accountability, and transparency (FAT), with trust often construed as an overarching aim that these contribute toward. This relationship is perhaps most clearly represented by the guidelines produced by The European Commission’s High-Level Expert Group on AI, which identifies fairness, accountability, and transparency as three of its seven key requirements for “Trustworthy AI” [4]. Similarly, the Montréal Declaration for Responsible AI states the collective impact of realizing its enumerated principles (which allude to FAT though adopting slightly different terminology) as “lay[ing] the foundation for cultivating social trust toward artificially intelligent systems” [29]—their accompanying report [28] using the word “trust” over 40 times. In their review of documents containing ethical principles for AI, Jobin et al. [54] identify trust as one of the 11 common principles (featured in 28 out of 84 documents) alongside transparency, fairness, and responsibility/accountability; and specifically found that 12 of these documents viewed transparency as key to fostering trust.

Despite consensus that Trust in Artificial Intelligence (TAI) matters, precious little research has focused specifically on TAI; and the little that does exist typically does not do justice to trust. Research typically assumes that the technology which trust relates to is complex, but trust itself is not; that the term “AI” unhelpfully collapses a wide range of disparate systems, but the term “trust” is understood the same by all. Consequently, trust often remains under- or even altogether undefined, and its dimensions and complexities are overlooked. Fundamental questions remain regarding “what trust is, and how it can be built and maintained” [131]. This paper endeavours to begin closing this gap.

In what follows, we explicate the myriad facets of TAI explored across the literature. In section 2, we analyze an established corpus of TAI publications and uncollapse the manifold ways in which trust and distrust affect the perception and use of AI. The resulting ontology reveals the full richness and import of TAI, highlighting synergies and tensions between various orientations. In section 3, we demonstrate the usefulness of the ontology by reflecting it through a second corpus of FAT publications, using it to expose under-developed dimensions of TAI and diagnose critical disconnects in how researchers justify their TAI efforts.

Contributions:
(1) We develop a common frame of reference for explicating different approaches to the multifaceted challenge of TAI to promote greater clarity in the discourse.
(2) We identify core pillars representing prominent and relatively coherent TAI narratives that may be useful for orienting one’s own approach to TAI.
(3) We demonstrate the further usefulness of this common frame by applying it to FAT literature to reveal important differences in conceptions of TAI underlying the three dominant approaches to trust-promotion, as well as potential anomalies deserving of attention.

2 SURVEYING THE FIELD
The first corpus of texts underlying our review of trust in AI literature was created through a moderated process of selection and group discussion. The working group tasked with compiling a list of texts relevant to the topic of trust in AI was assembled by the Partnership on AI (PAI) (see the acknowledgements section at the end of this paper). It consisted of 30 industry professionals and academics who had significant expertise relevant to the topic of trust in AI, with a more active core group consisting of 10 individuals including the authors of this paper. Working group members

1While the conclusions about the state of the literature to date presented in what follows owe a number of impulsive to the discussions of the working group, as summarized in a Key Insights document (available: https://tinyurl.com/3pwfjpxs), here the authors have tried to go much further in-depth analysis regarding axes of difference within the literature.
were asked to submit their suggestions for relevant articles to an online repository, which was subsequently shared with the group. Members were then asked to highlight texts on whose inclusion in the repository they disagreed; texts were removed if at least two working group members had recommended this independently of each other. After the removal of these texts as well as duplicates, the repository comprised 78 texts, mostly academic in nature, but also including a number of industry reports and some items on the topic which were slightly more journalistic in style. Regular catch-ups and discussions were scheduled for the working group via Zoom. Initial discussions focused on understanding the nature and composition of the corpus of texts assembled. One important topic of discussion was apparent bias within the corpus towards individual concepts (notably: explainability) and whether or not this bias should be corrected. It was decided to expand the corpus further in order to fill in several blind spots which had been identified in group discussions (e.g. conceptual treatments of trust stemming from social science).  

The authors of this paper were given the task of assigning codes to the papers in this first corpus, principally based on their abstracts, which were designed to capture whether a paper was motivated by concerns relating to what trust means in the context of AI, including different dimensions of trust (a category we named “understanding”; 78% of the corpus); the specificities of different objects of trust in this field (“receiving”; e.g., trust in data vs. trust in an algorithm vs. trust in the company owning and deploying the algorithm; 37%); the different antecedents of and/or interventions affecting trust or distrust (“promoting”; 64%); and/or the different consequences that trust and distrust may have (“impacting”; 16%). While these codes were developed and assigned for a purpose separate from this paper—an annotated bibliography has been made available by the Partnership on AI 3—the exercise revealed challenges in unifying a diverse set of social phenomena referred to as ‘trust’ [14, 34, 94, 95, 119, 120, 136] and the casual appropriation of these myriad phenomena within TAI.

In order to more formally account for the differences we were observing, we carried out a two-stage qualitative content analysis [50]. Emulating conventional content analysis, stage one developed categories through a process of induction, resulting in key axes of difference organized around the questions: Who is doing the trusting? …in what? …on the basis of what? …in order to what? …and why? Categories (i.e. inferred responses to each of these questions) were arranged into an initial web structure, which was used to guide the directed content analysis of the next stage. Stage 2 used a directed content analysis approach [50]; papers’ orientation to “trust in AI” was analyzed with respect to the agent doing the trusting, the object of the trust, the basis for trusting, the objective (e.g. to gain, maintain, or restore trust), and the proposed impact of having that trust, corresponding to the organizing questions above.  

The aim of this exercise was three-fold: 1) to identify exemplars of each of these categories, 2) to validate and/or extend this emerging ontology to reflect the full diversity comprised by the corpus, and 3) to look for relationships between categories, specifically whether there were identifiable pathways through the diagram which might represent shared orientations within the corpus. The findings of stage 2 are reported below; the final iteration of the ontology is depicted in Figure 1, which for completeness includes an additional category, promotion, i.e. the proposed mechanisms for promoting trust (see section 2.6).

### 2.1 Agent: Who is doing the (dis)trusting?

We identify three broad types of trustor. The first (and dominant category) are the users who knowingly interact with AI during the accomplishment of a task, such as decision making [19, 79, 99, 115, 116]. These interactions raise a variety of concerns helpfully taxonomized as “use, misuse, disuse and abuse” of the AI [106]. Explorations of use reveal ongoing “re-negotiation” of trust in the AI “through practices of skepticism, assessment, and credibility” [109]. Misuse involves the use of otherwise trustworthy AI in ways that make its effects untrustworthy, in some cases leading to catastrophic failures and complete erosion of trust (see [68]). Misuse can result from abuse, i.e. implementing AI “without regard for the consequences for human performance” [106], and can take the form of “neglect” [43]. Often misuse occurs as a result of over-trusting [142] or over-relying on [19, 35, 75] the AI. Interestingly, publications in the corpus present evidence of *automation bias*, whereby people tend to defer to machines in instances of doubt over conflicting information [142], as well as the contradictory impulse of *algorithm aversion*, whereby “people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake” [30]. Explanations for this seeming incongruity include differences in agents’ predispositions (including both trusting stance [43, 95, 135] and risk tolerance [34]); membership in an affected group [146]; and level of AI expertise, competence, or attentional capacity [43, 125]. Other literature is concerned with the potential for agents to disuse the AI if believed to be insufficiently trustworthy [86], with one means of preventing this being to assist users in predicting when a system might fail [99, 115] (consistent with strategies for maintaining trust between stakeholders [62, 64]). Some of the literature suggests that when the AI behaves in ways that do not meet users expectations, some forms of transparency can mitigate loss of trust [61] and empower more effective use of these systems [112].

In contrast to the above, the subject is generally not using AI with intentionality, nor are they in a position to disuse it. Rather, this agent is affected by the AI’s recommendations [16] or wider impacts on the society in which they live. Often this subject has limited understanding of AI, but may perceive the AI as generally untrustworthy and experience high levels of anxiety about its impacts [21]. These feelings may stem from subjects’ comparative lack of agency and disproportional benefits accruing to entities who own the AI [60]. Subjects may seek to redress this imbalance by restricting access to their personal data for use by these systems [147]. To understand and/or avoid unfavorable decisions being made about

---

2As a result of deliberately enhancing the corpus with theorizations on trust, not all papers mention AI.

3Interestingly for this corpus on “Trust in AI”, not all papers specifically mention trust. For these publications, the reason for its inclusion and how it speaks to the matter of trust was inferred. This is consistent with our chosen methodology, which is concerned with differences in meanings rather than the number of instances per category.

4We note the particularly wide spectrum of user types explored across the TAI literature—from non-expert to expert, interacting with the AI in low- to high-risk contexts. These differences matter and require careful consideration.
them by specific AIs, subjects are thought to be interested less in the ‘how’ of the AI and more the ‘why’ [141], but may face significant challenges in contesting these decisions [140].

Finally, the controller is concerned with assuring the trustworthiness of the AI, either as the entity developing [115], purchasing (see [49], omitted from corpus), monitoring/assessing [80], or regulating AI (alluded to in but not a particular focus of this corpus). It is important to recognize that these types of controllers can have quite different agendas and capabilities. We have grouped these controllers together because they all determine the characteristics of the AI in a substantial sense, and do so through the application of formalized trustworthiness criteria [32, 87]. Problematically, these may not map neatly to criteria subjects or users apply in evaluating AI’s trustworthiness. Lay notions of fairness, for example, are not always aligned with statistical fairness constructs [15, 16]; moreover, there is an important temporal dimension to fairness requiring ongoing assessment of impacts [87]. In addition to efforts to bring these into closer alignment [77, 79] (see “goal congruence” [125]), there is a recognition of a need to bridge the divide between regulators’ and independent auditors’ trustworthiness criteria and that of developers and purchasers [138].

2.2 Object: In what is the agent (dis)trusting?

Most often, the object of trust within TAI is the technology itself. Setting aside the enormous variety of AIs, there are different characteristics of the technology that can be (dis)trusted, including “(1) the performance of the technology, (2) its process/attributes, and (3) its purpose” [125]—roughly relating to how well it works from the perspective of the user [33, 126], how it works [112, 115], and what it does [8, 79]. A key route to trust in the technology is through affecting users’ perceptions of performance through model interpretability [25, 32, 85, 104]. Explanations can be provided to compensate for interpretability shortcomings [39, 71, 115, 148], though do not always lead to greater trust [27], e.g. if users are given too much information [61] or are not given the right kind of information [19, 84]. Explanations necessarily elide certain details about the model’s inner workings which may prove important for being able to appropriately calibrate trust [99]. Counterfactual explanations, which convey why one thing happened rather than another, have been proposed as more useful to enabling subjects to act, above and beyond (or without need of) understanding how the AI works [98, 141]. Understanding how the AI works can, however, be useful for correcting mistakes [71], and in some cases individuals may wish to modify the algorithm itself to enhance their trust in
it [31], or defer to humans regarding “human aspects of context” [12] (also [89]) when the logical outcome is not necessarily the (morally) “right” outcome [102]. Whether or not the technology is functionally trustworthy, as in it can be trusted to work as intended, questions of trust might arise with respect to the application of that technology (“its purpose”), e.g. the trustworthiness of AI for automating driving [144], for medical decision making [86], or indeed for any role formerly performed by human actors [76]. An especially important concern is whether the AI can be trusted to promote justice [16], and an emphasis on understanding “the mechanics of AI” may in fact distract from or confuse deliberations on “the purposes to which AI can be put” [46] (postdating corpus).

A number of publications suggest a complicated intermingling of trust in the technology and trust in the organization [136]. Trust in AI is not simply a proxy for trust in the people who developed the AI [92]; however, trust in AI depends in part or in whole on the perceived trustworthiness of the organization developing or deploying the AI [48], and the loss of trust in the algorithm has been found to lead to loss of trust in the organization behind it [146]. This suggests the applicability of models of trust originating in organizational studies (a number are included in this corpus [14, 67, 81, 119, 123, 137]), though the aforementioned intermingling of objects of trust necessitates caution in simply appropriating these without careful nuancing. The TAI literature shows that organizational dynamics, such as norms and expectations and relationships with coworkers (a.k.a. “environmental factors” relating to team collaboration [43]), can influence reliance on AI outputs [75]; and that the use of AI in organizational contexts also introduces dynamics of “meta-trust”, or “trust a person has that the other person’s trust in the automation is appropriate” [75]. While we believe trust in the organization is a useful category, we note that there may be quite important distinctions between types of organizations that influence the nature or quality of trust agents have in them. In particular, in contrast to private sector organizations developing or deploying AI, there is an expectation that public entities are obligated to protect individuals from harm [138]—particularly as the agent is typically a subject of public sector AI rather than a user.

Lastly, it can be structures that one trusts, or the wider institution comprised of these structures (see “institutional trust” [1, 95]).6 Structures relate to so-called “deterrence-based trust” [119], i.e. trust that laws and sanctions are sufficient to ensure AI’s trustworthiness.7 Public anxiety about AI is linked with a perception that those deploying AI (whether private or public entities) operate without restraint [22]; thus a key question of structures is the strength of oversight mechanisms [147] and whether AI can be held accountable to legislative controls placed on it [9, 12]. This “macroscopic societal accountability” is not satisfied through well-crafted explanations of single AIs [1], but rather through building trust in broader social systems. It may be an indicator of structures in need of trust-building that reliance on big data for decision making [60, 80] is more prominent in cultures where institutional trust has eroded (also see [109]); and that users (e.g. of Bitcoin) “prefer algorithmic authority to the authority of conventional institutions, which they see as untrustworthy” [89] (a finding supported by [122]).

2.3 Basis: Forming (dis)trust on the basis of what?

TAI largely prioritizes a cognitive view of trust that emphasizes understanding as the principal mechanism of trust-formation. Transparency, interpretability, and explanation are direct appeals to this cognitive mode; however, certain cognitive limitations may hinder these approaches. For one, these approaches can place unreasonable demands on the explainee’s time and energy [18, 39, 62, 71, 104]. Assuming high motivation to understand the information, however, agents may yet be limited in their ability to understand or make use of it [5, 112, 141, 148]—explanations are a social interaction, requiring that the explainer understands and responds to the beliefs and needs of the explainee [98]. And even for the most willing and able agent, it is not possible to see what AI does at all times [5].

Importantly, a “too narrowly cognitive” view also neglects “emotional and social influences on trust decisions” [67]. Social (“relational”) [119] or “interpersonal” [81] trust arises through interactions which reveal the trustworthiness of actors implicated in the development, deployment, or use of the AI. For example, people have been found to engage with Facebook as a “quasi-person” on the basis of conceptually linked technology and interpersonal characteristics: “competence-functionality, integrity-reliability, and benevolence-helpfulness” [74]. Much of the TAI literature draws from the legacy of the well-known Ability-Benevolence-Integrity model [123] in articulating key concerns underlying social trust (e.g. [74, 111, 136]; also [134], not in corpus); but organizational dynamics can also affect trust, as subjective assessments of interpersonal reliability play out within the context in which the AI is deployed [43, 75] (see also reflections on agency in fostering trust [26, 66]). Attending to the latter, TAI can draw from the fact that people behave more trustworthy when they know they are going to be held accountable [37, 64, 117]. Interestingly, Interactive Machine Learning offers an opportunity for users to form a more inter-personal relationship with the AI itself, with each party affecting the behavior of the other [33].

In addition to rational bases for trusting—whether cognitive or social reasons for believing in the object’s trustworthiness—there can be an affective dimension “grounded in reciprocated interpersonal care and concern” which moderates trust formation [193], not in corpus). This affective trust relates to the “emotional bond among all those who participate in the relationship” and the feelings associated with betrayal of trust [82]. While this could certainly apply to organizational objects of trust, it is an interesting question whether, or to what extent, this applies to the human-AI (technology) trust dynamic, as there is no presumption of social exchange in the technological relationship [75], nor the reasonable expectation that this bond will compel AI to act more trustworthy [82]. People are often thrust into relationships with AI which force them to build trust in ways they wouldn’t normally do socially [75], though susceptibility to trusting anthropomorphized AI reveals a desire for individuals to form a familiar social bond [78, 144] despite the inherent social limitations of AI [102]. Then again, this bias is greatly dependent on whether the agent believes humans are
inherently trustworthy as compared with technology [135], and in which situations [75, 76]. Still, the literature finds evidence that agents respond positively to perceptions that the AI is “honest” [126] and respects their autonomy [39]; and negatively when the system betrays their expectations and/or annoys them [39]; and representation can be altered to manipulate affective responses [125], as with anthropomorphization. When people have limited rational basis for trusting (e.g. lack of access to evidence of trustworthiness, or difficulty comprehending this evidence), the affective component may play a greater role in not making, or not, the leap to trust (see [100]). This is evident in the public’s aversion to “scary” robots [22] and fear of myriad consequences of AI [8]. When trusting in structures, perceptions of situational normality (see also “presentational base” [82]), or the general sense that “everything is in the proper order” [125], is conducive to trust.

## 2.4 Objective: What is the trust goal?

AI is a new addition to the cultural landscape which does not benefit (nor should it) from presumptions of trustworthiness generally afforded to more established social institutions (see [82]). Many agents, therefore, are in the initial trust-forming stage of their relationship with AI [1], and their first impressions of performance and process need to be positive to gain trust [125]. Gaining trust is also an objective as it relates to agents who, for whatever reason, begin with low levels of trust [48]. Typically, trust is gained slowly through direct experience of and continual reassessment of trustworthiness [81, 109, 123, 137]. While this process works well for users, who can respond to encounters with decisions and explanations provided about those decisions and re-calibrate trust accordingly, subjects are often not even aware of their interactions with AI and may only become aware of the AI when it negatively impacts them (see [65, postdating corpus]); though one proposed means of gaining subject trust is through public engagement (e.g. citizens’ juries) to elicit the public’s thoughts on ethical or unethical uses of AI [8].

Different mechanisms are involved in maintaining trust (see also “continuous trust” [125]). Maintaining trust involves continuing to nurture trust while simultaneously protecting against the loss of trust (particularly catastrophic loss of trust) [66]. Interpretability and explainability research tends to be in service of intervening within an interaction that might otherwise reduce a user’s trust (e.g. [71, 83, 84]). As noted previously, maintaining trust in the technology is particularly difficult given people’s intolerance of mistakes made by AI [30] and issues with users being able to sustain interest in explanations [18].

The objective of restoring trust when it is lost may yet again entail different tactics, principally because the agent’s stance may be characterized as distrust [14, 95]. For example, restoring subject trust in AI would involve active dismantling of culturally ingrained narratives and replacing them with more positive or neutral ones [21, 125]. Some scholars conceive of distrust as being a state of lower trust than that brought to initial encounters [123] (i.e. in comparison to the gain condition), meaning that greater effort is required to raise trust to the desired level. But distrust is more than just lack of trust; it is an active stance, making it much more difficult to recover from (see strand of literature on trust repair, e.g. [7], not included in corpus). Distrust often results in disuse, which tends to result in fewer subsequent trust-building experiences (see [14]). In addition, distrust tends to be associated with negative affect arising from trust having been violated [67], as well as the entrenchment of suspicion [67], and these are not easily resolved through provision of evidence of trustworthiness.

## 2.5 Impact: What are the benefits of having trust?

The logic underlying the epistemic rationale for the importance of trust in AI is that productive synergy between user and AI arises when trust aligns with correct beliefs regarding trustworthiness [19, 75, 142]. Helping the agent to form a correct (or at least interacionally beneficial [3]) mental model is key to ensuring an effective dynamic based on appropriate reliance [18, 35, 43, 85, 112, 115]. This involves both being able to control the AI (e.g. guiding it toward the correct decision, and making sense of the AI output [1], particularly in instances when the AI’s decision does not match that of the user. Experimental evidence shows that people trust the AI more and adjust their own predictions to match the AI’s if the model’s accuracy is observed to be high, but are generally unlikely to do so merely on the basis of high stated accuracy [148].

Many statements on the importance of trust in AI are instead couched in instrumental terms. It is commonly asserted that trust is critical for realizing the full economic and societal benefits afforded by AI, insofar as lack of trust will lead to lack of adoption across the economy [30, 34, 48, 76, 120, 125]. While this argument may hold strongly for purchasers (a type of controller), user and subject distrust may not be determinative of non-use [63] due to various social and/or organizational entanglements with the technology that prohibit disuse. Other instrumental TAI arguments include usage continuance [30, 31, 43, 61, 74, 84, 106, 111], permitting access to data for use within the AI [147], and promoting productivity and/or efficiency within organizations using AI [109, 137] (see also [64, 94, 117]).

Less commonly, trust in AI is seen to be intrinsically valuable. In sociological literature, trust is considered “a functional prerequisite for the possibility of society” [82], the foundation of all social relationships (see [136]), and it enables humans to take action on the basis of expectation despite the inherent unknowability of the future [100] (see also [66, 109]). Trust is not only an enabler of a functional society, it affords psychological benefits to its members, e.g. mitigating fear when facing vulnerability [67], feeling confident about one’s actions despite uncertainty [32, 86, 125], and experiencing an emotional bond [135]. The intrinsic view makes room for healthy distrust as a key to a harmonious relationship with AI: People may be willing to rely on AI while being aware of issues with the AI that pose a threat to this relationship (see [95]). Regardless of ultimate adoption, distrust tends to signal that “important matters are at stake” [63] and, if taken seriously, can exert a corrective force on the development of AI which brings it into closer alignment with human values (e.g. [8, 16, 79, 138, 146]). It may be that, given limited agency in disusing AI, not to mention the challenges in developing ethical AI [22], distrust may in fact be the most appropriate “mechanism for dealing with risk” [95] entailed by the technology.
2.6 Core pillars

Figure 1 illustrates the complexity of TAI and the many different ways it might be approached. Assuming any given orientation to TAI is comprised of one category of agent, object, basis, objective, and impact, and that none of these are inherently incompatible, there are potentially as many as 243 combinations, each representing a different facet of TAI. That is not even accounting for the many subcategories that have been abstracted away—noteably, different types of user (non-expert versus expert, in various contexts of use) or controller (developer, purchaser, monitor, regulator), and of course, a huge variety of types of technology. And yet, emerging from this ontology is a set of three fairly coherent core pillars within the TAI literature (see columns, bold vertical pathways). The first pillar, left on the diagram, is concerned with optimizing decision making: one must understand enough of what the AI is doing and its particular limitations to be able to determine what action to take given its output. Transparency is the intuitive means of promoting appropriate reliance so that the agent maintains trust in the AI as a tool to inform their decision making. The second pillar, middle, is concerned with maximizing uptake: organizations embedding AI into their products and processes must gain the confidence of those affected by the AI so that they embrace it/them. Commonly, this is done through efforts to visibly attend to the matter of fairness. And the third pillar, right, is concerned with minimizing harm: People must feel safe and comfortable in an increasingly AI-driven society. Widespread distrust of AI requires developing a system of structures, e.g. accountability mechanisms, that ensure that AI is trustworthy.

3 TREATMENTS OF TRUST IN FAT LITERATURE

In the 2+ years since the curation of the above corpus, there has emerged a sufficiently large, focused collection of works on fairness, accountability, and transparency to analyze for treatments of trust. To create our second corpus, we searched all papers to date appearing in the ACM Conference on Fairness, Accountability, and Transparency (formerly FAT∗, now FAccT) proceedings for instances of “trust” within the abstract, keywords or body of the paper. Papers were included if they used at least once the term “trust”, “distrust”, “trustworthiness”, and/or “trustability” (excluding uses of these terms only in list of references, and excluding the legal term “anti-trust” and the phrase “absolute trust” in statistical measures [132]); we also excluded papers that did not provide enough surrounding context when mentioning trust to derive any codes (only one paper: [58]), papers using trust only when describing study limitations [23], and papers where the only mentions of trust were in reference to something other than AI (e.g. a mass spectrometer [38]). This resulted a total of 57 papers, or 24.3% of the total FAT∗/FAccT proceedings—a very significant proportion, indicating the centrality of trust and justifying its thorough treatment here. Emulating the directed content analysis performed on the first corpus for this second corpus, we coded all papers according to agent, object, basis, objective, and impact. It is important to note that this was a non-trivial exercise. Codes needed to be inferred, as authors do not have a common language for talking about their particular orientation to trust, and indeed may not realize that their orientation may differ from others and thus need spelling out. This makes the ontology highly valuable, as it will enable authors to establish their position relative to a common frame.

Our analysis did not surface any additional categories unaccounted for in the ontology—itself an interesting finding given that this need not necessarily have been the case. There are in fact many similarities between the two corpora; in our summary of the FAT corpus below we focus on differences, highlighting new contributions from and particular emphases by the community. But beyond this, we hope to show that the ontology can be usefully applied to better understand individual papers and bodies of works. In section 3.6 we explore how the technical levers for promoting trust (i.e. fairness, accountability, and transparency) refract through this lens, and raise questions about potential anomalies.

3.1 Trusting agents in FAT literature

The FAT corpus contributes new (if limited and conflicting) data regarding potential human biases regarding AI. While one paper finds evidence of automation bias in high-impact decision making contexts, with students perceiving automated decision making (ADM) as fairer than human decision making (HDM) for university admissions [91], another finds a lack of consensus whether ADM was fairer than a human judge for decisions relating to granting bail [47]. The corpus also shows continued interest in (in)appropriate reliance [149], with concern regarding users’ susceptibility to disinformation [103], as well as new findings that although users can “somewhat differentiate correct machine predictions from incorrect ones,” these instincts can be deceived by plausible yet spurious explanations [73]. While some work focuses on trust by users who are highly motivated to inspect the AI’s trustworthiness [20, 105], the dominant concern of corpus 1, other work questions the interest users have to engage with the AI at all (e.g. doctors preferring to be patient-oriented [124]) and the capacity of subjects to engage in meaningful deliberations on AI’s trustworthiness [65].

There is a particularly strong focus in this corpus on the trust of the subject, including a detailed examination of public TAI as qualitatively different to TAI of users [65]. Some works allude to users’ dual role as subject of the effects of a data system [6, 129]. The focus on subjects is further evident in works exploring the challenges of crafting subject-actionable explanations, contributions deepening understanding of when to use as well as how to create counterfactual explanations [11, 99, 101] and offering other methods to facilitate algorithmic recourse [56, 139]. Compared to corpus 1, this corpus also shows a growing interest in the trust needs of controllers, with new methods to help guide developers in developing and evaluating trustworthy AI [128] and communicating trustworthiness to regulators [127].

A major contribution of FAT publications is an emphasis on trust being particular to a given agent and the challenges in both balancing [90] and tailoring trust-promoting efforts for multi-user

---

8 In reality, it is common for papers to speak to multiple categories at once when exploring TAI.

9 We recognize that other venues might have been included which have relevant contributions to FAT literature. Our analysis to follow is limited to the ACM FAccT community during the years 2018–2021. Future work could apply this ontology to compare FAT venues and/or look for changes to FAccT over time.

10 Two of these papers overlap between corpus 1 and corpus 2: [79, 99].
[42, 133] and multi-stakeholder contexts [124, 130, 143]. Works illustrate tensions between the user's needs (e.g. to understand model performance) or controllers' needs (e.g. to evaluate model fairness) and the needs of the subject (e.g. right to privacy) [6, 17, 107, 130]; the importance of attending to subjects' diverse needs, with evidence of clear hierarchies of beneficiaries [143]; and the limitations in usefulness of explanations to users compared to controllers [13]. This corpus clearly responds to the call of corpus 1 to reconcile developers' and regulators' assessments of trustworthiness [65, 70, 90, 127, 128]. While there is a growing interest in audits as a means of promoting trust [65, 118, 145], this corpus reiterates the concern that fairness constructs differ between agents [57, 90, 127, 145]—e.g. what a controller would consider 'unbiased' is not the same as what a subject would consider 'fair' [47]. There is, furthermore, a concern that AI audits may promote unwarranted trust, with reliance on statistical fairness metrics flattening the discourse and discouraging deeper analysis of structural inequality [24, 40, 72].

3.2 Objects of trust in FAT literature

The majority of transparency and fairness contributions within the FAT literature are concerned with trust in the technology itself (see Figure 2). This is accompanied, however, by critique that a focus on trust in the technology perpetuates the problematic notion of model objectivity [40]; that, indeed, “trust in a technology is rooted in relationships—not in a technical specification or feature” [124]. We also see a further blending of trust in technology and trust in the organization, as seen with concern regarding management of untrustworthy data sources [103]; discussion of how AI narrows socially constructed categories [45, 97]; and explorations of the trustworthiness of the human processes involved throughout the AI development pipeline [113], including in problem formulation [108], model development [51], data labeling [38], and creation of explanations [44].

A resounding call from the FAT corpus is for more situated understandings of trust within deployment contexts [2, 10, 124] and even “re-contextualising data and models” [121]. Technology as not a thing one can trust absent context, as it may be trustworthy (e.g. ‘fair’) in one context but untrustworthy in others [118, 127]; thus, new methods are needed for making boundary conditions interpretable [127]. There is, further, an emphasis on the role of “human infrastructures” [121] in facilitating dialogue that enriches understanding of context in ways that both improves trustworthiness and promotes interpersonal trust in organizations (see also [59]). Perhaps most notable of this corpus is the discussion of the effect of power dynamics on (dis)trust in a given a context [2, 10, 24, 72, 121]. Works emphasize the importance of empowering subjects as beneficiaries of these data systems to mitigate distrust of organizations [2, 121], and fundamentally challenging the structural contributors to decisional power asymmetries [10, 59, 72].

This corpus demonstrates rising interest in promoting trust in structures to compensate for low levels of trust in organizations’ commitment to mitigating AI harms [72]. Efforts include promoting accountability for AI systems [55, 70, 113, 118]; establishing clear processes for subjects to challenge unfair decisions [56]; establishing robust intra-organizational governance practices [51, 65, 113, 145]; and empowering a network of controllers (e.g. regulators, auditors, journalists, advocates) [65, 96, 114, 118, 121, 145]. Interestingly, a number of publications characterize the technology itself as structure, insofar as the AI can contribute to systemic bias [40, 130] and growing power imbalances [2] which contribute to distrust. The intermingling of these different objects of trust is evident in the effort made within the FAT corpus to attend in a more rounded way to trust in technology, organizations, and structures [2, 121, 124]—in short, cultivating trust “ecosystems” [65, 121] that promote trustworthy technology and impel trustworthy organizations.

3.3 Bases of trust in FAT literature

As with the first corpus, a cognitive view of trust dominates the transparency, interpretability, and explainability contributions, at least within the FAccT conference archives. And yet, these works reflect waning confidence in the usefulness of explanations to end-users—non-expert end-users rarely trust the explanations [44], decision-making by expert end-users is not improved greatly by explanations [53], and contrastive explanations that succeed in helping users understand prediction errors nonetheless do not improve trust in the model [88]. The community calls for the development of consistent criteria for evaluating explanations [128], with resultant trust being a key consideration. The corpus clearly acknowledges a social (i.e. interactional) component to successful explanations, with works exploring how to make explanations more intuitive so that they ring true [110], tailoring explanations to the informational needs of recipients [128] (including preserving subject autonomy [11]), and even asking whether explanation recipients can contribute to constructing explanations that work for them [45].

Notably, the FAT corpus places slightly more emphasis on social trust than cognitive trust, and proportionally much greater overall emphasis on social aspects, perhaps reflecting the social science leanings within the FAT community. Two in-depth explorations of TAI within the corpus [52, 134] draw heavily from social trust models—both seeing TAI as “rooted in, but nevertheless not the same as, interpersonal trust as defined by sociologists” [52]. The corpus raises a number of new social considerations. One is that subjects’ belief in an organization’s cultural dissimilarity can reduce trust [2]. Not only does dissimilarity breed suspicion regarding the organization’s Benevolence [2] (as in the Ability-Benevolence-Integrity model), it may affect their Ability to deliver TAI. The work highlights the importance of embracing cultural differences in notions of fairness [2, 47, 121] (see also [91]); taking into account the on-the-ground social realities of users and subjects [2, 41, 59]; co-developing measures of risk with affected communities so that they reflect real harms people experience [59, 72, 96]; and being seen to make good faith efforts to balance different stakeholder interests [90, 91].

The affective dimension remains comparatively under-developed, despite allusions to unaccounted for dynamics limiting trust users and subjects have in both the technology and the organization. The lack of structural guarantees of trustworthiness has been identified as creating conditions for affective distrust to thrive [65], as is the sense that one has no recourse to algorithmic harm [139] and that organizations are all-powerful [55, 143]. A proposed counterweight is obligating organizations to act in the best interest of
those affected by their AI, as in a fiduciary model [11]. The corpus is also greatly concerned with subjects experiencing harm (more on this in section 3.5), demonstrating an understanding that feeling betrayed or insufficiently cared for diminishes trust in ways not easily recovered from [2].

### 3.4 Trust objective in FAT literature

*Maintaining* trust is less robustly linked with the transparency, interpretability, and explainability contributions in the FAT corpus compared to the first corpus. This may be a function of the particular kinds of agents these corpora focus on. *Maintaining* trust is a key objective in relation to agents whose use is required and/or whose trust facilitates improved decision making outcomes [53, 73, 149]; whereas *gaining* trust is more salient to those whose use is in doubt and/or whose trust is important for intrinsic reasons [13, 78, 88, 105].

By far the most common objective within the FAT corpus is *gaining* trust; however, the literature reiterates the importance of cultivating understanding of when to trust the technology, proposing the terminology of “calibrating” trust as more apt [149]. Indeed, a particularly noticeable feature of the FAT corpus is its exploration not of how fairness, accountability, and transparency promote trust, but how these either can or should promote distrust. It is noted that highly usable visualization tools promote trust in the technology by data scientists without requiring them to understand how it works [45]. There is a concern that efforts to improve representativeness within datasets can inspire trust, lending the AI “rhetorical authority, merited or not” [24]. It is also argued that AI may promote unwarranted trust in both organizations and structures, as it “plays a particularly important role in political contexts rife with distrust, in which officials facing external scrutiny need to depoliticize their actions by ‘making decisions without seeming to decide’” [40]. Similarly, China’s Social Credit System attempts to compensate for high levels of distrust in society through dubious attempts to quantify citizens’ trustworthiness [36].

*Restoring* trust is a more common objective in this corpus than the first, perhaps resulting from greater attention to the matter of distrust. Two important contributions emerge. The first is that public distrust in organizations is seen as contributing to tighter controls over access to demographic data, which limit controllers’ ability to evaluate the fairness of algorithms [6]. The second is to shine a spotlight on the need for particular care in *restoring* trust in contexts where distrust results from legacies of colonialism, disempowerment, and/or discrimination [2, 10, 17, 40, 107, 121, 130, 143].

### 3.5 Impact of having trust in FAT literature

The various FAT publications concerned with enabling agents to identify when the technology is untrustworthy [20, 73, 127, 149] are clearly justified by the *epistemic* rationale for TAI: Knowing when to trust AI is key to using it in ways that lead to improved decision making (see also [20, 56, 99]). Interestingly, this argument is extended to collaborating data scientists who need to know which datasets and algorithms produced by which collaborators are trustworthy in order to better inform policymaking [133]. A new, and particularly striking contribution of the corpus, however, is its problematization of trustworthiness. Works call for deeper examination of the “epistemic and methodological underpinnings of algorithms” [40] (italics added; see also [24]), questioning whether algorithms are faithful enough representations of complex phenomena to derive trustworthy outputs [45, 108] and whether explanations “reflect real patterns in the data or the world” [44], how to determine when datasets are representative in a meaningful way [24], and whether TAI principles can be implemented in a way that can be audited [90].

As with the TAI corpus, we see the *instrumental* rationale in (limited) works concerned with adoption [88], as well as those noting that subjects’ distrust of organizations presents a barrier to collecting data [6, 121]. As to the latter, attending to algorithmic fairness is thus seen as *instrumentally* valuable in “enabling data collection in low-trust environments” [6]; whereas other works propose fairness as important for reputational preservation, including preserving trust in institutions (e.g. higher education) [91]. There remains an implicit concern that distrust will limit uptake of AI. To address this, AI audits have been proposed as a means of engendering trust, but now the FAT community is wrestling with how to instill trust in the audits themselves [145].

Intrinsic rationales for TAI are much more prominent in this corpus than in the first corpus, perhaps because subject trust, a strong focus in FAT literature, is not easily justified in *instrumental* or *epistemic* terms. It is clearly seen as “fundamental to human flourishing” that people have “peace of mind” [139] regarding the use of AI in society. FAT literature emphasizes that this trust will only be earned by effectively mitigating AI harms [41, 59, 72, 90] (i.e. real harms to real people [96]), while at the same time ensuring more equitable distribution of AI benefits [143]. While it is common for TAI to frame distrust as largely unwarranted (e.g. see [21] from the first corpus)—seeing this distrust as possibly arising from a failure to communicate the technology’s trustworthiness—the FAT literature instead proposes distrust as the most appropriate stance towards AI at this time, given that potential harms and inequities have not been satisfactorily resolved.

### 3.6 A bird’s-eye view on FAT

Figure 2—breaking out the three promotion mechanisms individually—tells an interesting story about the FAT corpus. There are clearly very different orientations to trust between the transparency subset and fairness and accountability ones. And for the most part, these transparency publications conform to expectations, following down the left column the narrative of *optimizing decision making*. It is surprising, however, that here transparency is presented as a mechanism for *gaining* trust more than for *maintaining* it, particularly given that the TAI literature suggests that explanations are generally sought in moments when trustworthiness is suddenly thrown into question (i.e. when the AI does something unexpected).

The fairness literature also largely conforms to expectations, in line with the *maximizing uptake* narrative (middle column), though it seems odd that a predominant object of trust in these papers is
technology (10 of 23 papers). Even if most often papers focus on subject trust in an organization, and some of the fairness papers which explore trust in technology are concerned about user or controller trust, we have to question whether it makes sense to discuss subject trust in technology. Trustworthiness of the AI itself is largely beyond the capability of subjects to gauge, and often they are not even in a position to question the trustworthiness of the AI because it acts upon them in ways unseen [65]. Because of this, we might also have expected more of a recognition of the trust subjects place in structures as a means of ensuring fairness, i.e. freeing them from having to concern themselves with the AI’s trustworthiness [65]. It also seems odd that while trust is seen as instrumentally useful and the most common objective in the fairness works is to gain trust (see bold boxes), these are rhetorically disconnected within individual publications, the most common connection being between gain and epistemic (see bold line).

The accountability subset diverges noticeably from the expected narrative (i.e. not proceeding down the right column), leaving us wondering if there is a coherent justification for accountability other than minimizing harm. Granted, the accountability subset was the smallest of the three and perhaps more data are needed to see meaningful patterns; but why isn’t the controller a more important agent in these publications? Even among the ones looking at controllers’ trust, there is surprisingly little discussion in the accountability papers (at least the ones that mention “trust” that were included in this corpus) of trust by regulators. This is out of step with calls within the TAI literature for the need not only to regulate AI but, in order to do this effectively, to focus on what trustworthiness criteria regulators would look to (see section 2.1). For that matter, where is the focus on structures? It seems odd for controllers (if talking about regulators, that is) to trust in organizations, particularly on the basis of social trust, as opposed to trusting in an organization’s internal governance structures, for example their practices in producing algorithmic documentation [113]. Likewise, surely accountability mechanisms provide assurances of trustworthiness, and these assurances, the fact that it is governed by some capable entity, is what subjects would trust [65]. While we cannot answer these questions based on the current analysis, they may perhaps point to some of the confusions that make meaningful interpretation of ‘trust’ difficult, and hence, to areas where the ontology proposed in this paper could help unify the narrative going forward.

4 CONCLUSION

In developing an ontology of TAI, it is tempting to consider whether it is a true characterization of the literature. This, of course, is hard to say with certainty. The more interesting question, though, is whether it is useful characterization. So what does this model do for us?

The first thing it may do is reveal important distinctions that generally go unstated and therefore unacknowledged. Trust may be different enough between the myriad pathways through the model that it makes little sense to compare findings between papers following very different paths. Going forward, having a set of commonly understood referents may allow authors to make their orientations clear, helping to avoid issues of papers talking past each other and potentially muddling insights into how best (or even whether it makes sense) to promote trust in AI for a particular set of agents in a particular context.

At the same time, despite its complexity, the ontology reveals strong connections between certain entities, helping to simplify what otherwise may seem like disjointed discourse. The three pillars are recognizable, if necessarily simplified, ideal–typical categories within TAI. While clearly not all TAI contributions fit within these three pillars, they may be described in contradistinction to these ideal types as a way of progressing the field.
And finally, we have shown that the ontology can be usefully applied as a lens. We were able to more clearly see what aspects of TAI the FAT community has been most interested in and where their contributions lie; see how contributions on fairness, accountability, and transparency differ in terms of their conceptualizations of trust; and even challenge whether certain arguments are coherent and certain mechanisms for promoting trust are appropriately conceived.

5 ACKNOWLEDGMENTS
This work is partially funded by the ESRC funded grant BIAS: Responsible AI for Labour Market Equality (ES/T012382/1) and by the Data Science Institute at Lancaster University. The authors would like to thank the Partnership on AI for convening its expert group on Trust in AI, from which this collaboration sprang.

REFERENCES
[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhal. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In Proceedings of the 2018 CHI conference on human factors in computing systems. ACM, 582.
[2] Reidst Abebe, Kehinde Ataluvie, Abieba Bikehane, Sara Kingsley, George Ohade, Sekou L. Remy, and Swathi Sadugopa. 2021. Narratives and counternarratives on data sharing in Africa. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 329–341.
[3] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: and the limits of fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 648–657.
[4] Gregory A Bigley and Jone L Pearce. 1998. Straining for shared meaning in organization science. Problems of trust and distrust. Academy of management review 23, 3 (1998), 405–421.
[5] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency. PMLR, 149–159.
[6] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jian Zhao, and Nigel Shadbolt. 2018. ‘It’s Reducing a Human Being to a Percentage’: Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI ’18). ACM, New York, NY, USA, Article 377, 14 pages. https://doi.org/10.1145/3173574.3173951
The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency

[136] Daniel Ullman and Bertram F Malle. 2018. What does it mean to trust a robot?: Steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 263–264.

[137] Bart S Vanneste, Phanish Puranam, and Tobias Kretschmer. 2014. Trust over time in exchange relationships: Meta-analysis and theory. *Strategic Management Journal* 35, 12 (2014), 1891–1902.

[138] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI ’18). ACM, New York, NY, USA, Article 440, 14 pages. https://doi.org/10.1145/3173574.3174014

[139] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 284–293.

[140] Sandra Wachter. 2018. The GDPR and the Internet of Things: a three-step transparency model. *Law, Innovation and Technology* 10, 2 (2018), 266–294.

[141] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[142] Alan R. Wagner, Jason Borenstein, and Ayanna Howard. 2018. Overtrust in the Robotic Age. *Commun. ACM* 61, 9 (Aug. 2018), 22–24. https://doi.org/10.1145/3241365

[143] Anne L Washington and Rachel Kuo. 2020. Whose side are ethics codes on? power, responsibility and the social good. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 230–240.

[144] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (2014), 113–117.

[145] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Staary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[146] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI ’18). ACM, New York, NY, USA, Article 656, 14 pages. https://doi.org/10.1145/3173574.3174230

[147] J Patrick Woolley. 2019. Trust and Justice in Big Data Analytics: Bringing the Philosophical Literature on Trust to Bear on the Ethics of Consent. *Philosophy & Technology* 32, 1 (2019), 111–134.

[148] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 279.

[149] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.