# Honors Thesis Outline, Draft 2

Stanford University Dept. of Philosophy

Ryan Othniel Kearns

30 November 2021

## Contents

# 1 Introduction

## 1.1 Trust as an End in Explainable Artificial Intelligence

> Cooperation between agents, in this case algorithms and humans, depends on trust. If humans are to accept algorithmic prescriptions, they need to trust them.

This is a quote from the Wikipedia entry for "Explainable artificial intelligence" [13]. Plenty of researchers in the field of Explainable Artificial Intelligence (henceforth XAI) take *trust* to be a primary end goal for their research. To the layperson this sounds correct – as AI systems become increasingly complex and capable of superhuman tasks, it seems important that we learn to develop and evaluate trusting relationships with these systems. Trust is one dominant attitude determining the extent, productivity, and quality of our interactions with artificially intelligent systems. Moreover, AI systems are frequently adopted in modern healthcare, finance, security, military, and other contexts. A future in which trust is absent in human-machine relationships sounds increasingly dim as machines come to govern more and more of these important domains of our lives, especially outside of our control.

However, the operative terminology in XAI does not include trust, but instead concepts like explainability, interpretability, fairness, and robustness. Instead of reaching for our goal of trust directly, XAI researchers take these other four qualities as (strongly) *entailing* or (weakly) *prompting* trust(worthy qualities), in some way. Particularly the first two concepts, explainability and interpretability, are taken to be diagnostic criteria or "intermediate goals" for trust [2]. Researchers like Ribiero et al., 2016 have directly linked explainability to trust:

> Whether humans are directly using machine learning classifiers as tools, or are deploying models within other products, a vital concern remains: *if the users do not trust a model or a prediction, they will not use it.* [9]
>
> Emphasis in the original.

Moreover, other researchers have directly tied transparency to trust [8, 5, 4]:

> Communication with the user is essential to a successful interactive system - it improves *transparency*, which has been identified as a major factor in establishing user *trust* in adaptive agents... [4]
>
> Kim 2015, p. 102. Emphasis mine.

## 1.2 The Place for Philosophical Investigation

According to [2], XAI researchers turn to explanation and transparency as diagnostic criteria because trust is “hard to formalize and quantify” (210). This difficulty is worrisome – if we do not know what trust *is*, how can we take seriously any claims that explainability and transparency guarantee it? This is precisely where modern philosophical theories on trust could provide utility.

The predominant philosophical literature on trust for the past 40 years has focused on *interpersonal* trust, that is, trust relationships between two people. In the last decade, however, a growing number of philosophers have begun investigating trust between humans and non-human entities [6, 12, 1] and between non-human entities or “artificial agents” themselves [11, 7]. For now we call these cases of *impersonal trust.* Trusting AI systems is a particular kind of impersonal trust that deserves further conceptual clarity. In particular, I am not aware of any theories that handle explanation and transparency as important ingredients to trusting relationships, impersonal or otherwise.

## 1.3 Approach

Before we’re to assess the (explanatory, transparent, etc.) conditions under which trust is appropriate, we need a conceptual picture of impersonal trust that accomodates the roles of explanation and transparency. The purpose of this paper is to provide the groundwork for such a picture. I do this by surveying three promising theories of impersonal trust – Walker, Buechner, and Tavani’s “diffuse default zone” theory [12, 1], Nguyen’s “unquestioning attitude” theory [6], and Taddeo and Primiero’s “e-trust” theory [11, 7]. Then, I propose my own formulation that accommodates the best of these views while also making room for explanation and transparency. I do this *without* opining on the specific roles that explanation and transparency play for impersonal trust – that is left for future work – and instead just making room for such a role within the theory.

My argument is something like the following: first, I’d like to make the case for reducing notions of *trusting* and *trustworthiness* to quantification over specific trusting *instances*. A trusting instance concerns a particular trustor, a particular trustee, a particular action for the trustee to take, and a particular set of contextual factors, defined as a set of beliefs the trustor holds about the trustee. All three of the theories we consider acknowledge contextual factors as important for specific instances of trusting, yet none are prepared to handle them in a sufficiently formal way. I show that with the right formalization, we can accommodate explanation and transparency as operations on the set of contextual factors, able to move a trustor-trustee-action pair from an untrustworthy relationship to a trustworthy one under the right conditions.

# 2 Existing Theories of Impersonal Trust in Philosophy

## 2.1 "Diffuse Default Zones": The Walker-Buechner-Tavani Theory

Jeff Buechner and Herman T. Tavani are philosophers. In their 2011 paper "Trust and multi-agent systems," they defend Margaret Urban Walker's conceptions of zones of default and diffuse default trust, developed in Walker's book *Moral Repair.* Buechner and Tavani argue that experiments on commitment and trust within multi-artificial-agent systems should be instructive to ethicists under this diffuse default model of trust – in other words, that philosophers can learn from experiments involving artificial agents as well as genuine ones.

According to Buechner and Tavani, the following conditions pertain to a trusting relationship between $A$ and $B$:

1. $A$ has a *normative expectation* that $B$ will do such-and-such;
2. $B$ is responsible for what it is that $A$ normatively expects her to do;
3. $A$ has the disposition to normatively expect that $B$ will do such-and-such responsibly;
4. $A$'s normative expectation that $B$ will do such-and-such can be mistaken;
5. Subsequent to the satisfaction of the above conditions, $A$ develops a disposition to trust $B$. [1]

The authors call these conditions the "Strawson-Holton-Walker" model of trust, relying on contributions from philosophers Peter Strawson, Richard Holton, and Margaret Urban Walker [10, 3, 12]. Buechner and Tavani argue that notions of "normative expectation" and "responsibility" make this theory difficult to extend to cases involving artificial agents, to which we cannot simply ascribe responsibility. They incorporate another view, also from Walker, in which artificial agents are "part of the environment," and people come into trusting relationships with that environment.

Margaret Urban Walker's account of trust begins with the observation that trust is contextual and localized in space; in particular, there are places we feel safe and operate with varying minimum "default" levels of trust. Walker's examples often entertain large communities where such "trust zones" occur, like entire cities where we expect adherence to traffic laws from one another, including pedestrian foot traffic "laws" that aren't enforced (e.g., making way on the subway platform). We take such actions to be the responsibility of people within the community, and it seems that complete strangers can have these symmetric and reversible normative expectations of each other. Walker considers this a form of trust that is an "unreflective and habitual background" in many scenarios we find ourselves in.

Because zones of default trust may contain people who rarely or never meet in person, and yet nonetheless trust one another, the explicit individuals in the trust relation need not be spelled out in advance. Instead, Buechner and Tavani propose the notion of a "generic individual" partaking in the zone of default trust, which may be surrogate for a real individual, a group, or a non-human agent like a computer network [1] (43). It is this final observation that makes the Walker-Buechner-Tavani model a model for impersonal trust.

Walker, herself, considers a variation of this idea called "*diffuse* default trust" [12] (85). Her categorical example involves feeling resentful towards an airline for poor service, say after a day full of delayed and cancelled flights. When we feel let down by an entire airline, says Walker, it's not that we were relying $X$ to do $A$, $Y$ to do $B$, etc., but rather the airline itself, which is a "mode of organization that is supposed to... enable whatever individuals are filling organizational roles" (85). It's unacceptable to reduce this down to trust of each airline employee, since we very well say and mean that we trust things like airlines, and this deserves a semantic account distinct from the state that would be trusting each airline employee. The upshot here is that the organizational or operational mode that is the airline is a recipient of trust under Walker's diffuse, default model, and Buechner and Tavani make clear that things like networks of artificial persons can likewise be recipients of trust in this way.

It is worth noting that neither Walker nor Buechner and Tavani apply the diffuse default model explicitly to the case of explaining or trusting technology. Buechner and Tavani's intentions for promoting the view have to do with refining our ethical account of trust in interpersonal relationships, aided by experiments involving artificial agents. The fact that such artificial agents can be spoken of as trusting and trustworthy under this view, however, is what makes the model relevant for our discussion.

## 2.2 "Unquestioning Attitudes": The Nguyen Theory

In "Trust as an Unquestioning Attitude", C. Thi Nguyen establishes a theory of trust to accompany philosophy's typically agent-oriented theories. Most previous theories of trust hold only between agents, involve central requirements like goodwill, responsiveness, or reliance on commitments, and are marked by the possibility of betrayal.

Nguyen, by contrast, establishes a theory of trust based on an "unquestioning attitude" – by this theory, to trust something is to have a strong disposition to suspend deliberation about that thing. Nguyen defines: to "trust $X$ to $P$" is to:

1. be first-order disposed to immediately accept that $X$ will $P$, and
2. to be second-order disposed to deflect questions about whether $X$ will $P$.

For example, a climber trusts their climbing rope (to hold them), because they aren't constantly reassuring themselves as to the rope's integrity. By contrast, to *distrust* is to remain in a state of constant questioning and skepticism.

Nguyen claims that both theories of trust – the general agent-oriented theory and the unquestioning attitude theory – sit under an umbrella for a more general notion of trust, and that what joins them is their purpose of expanding one's agency through integrating aspects of the external world. There is too much information in the world for one agent to account for all at once; therefore, we are required to trust and hold the unquestioning attitude towards some things. In this sense, we trust so to form linkages or "welds" to external objects, and we bring them into our practical and cognitive faculties. For this reason, we can talk of being *betrayed* in our trust even of non-agents, like musical instruments or the ground, because the "normative bite" comes from our desire to integrate objects in our immediate cognition and agency. When our climbing rope unexpectedly snaps, we feel betrayed because we had integrated that object into our agency and had developed an attitude of not questioning its integrity.

## 2.3 "E-Trust": The Taddeo-Primiero Theory

Philosopher Mariarosaria Taddeo, together with logician Giuseppe Primiero, develops a model of "e-trust", which is trust occurring in online or digital environments [11, 7]. Specifically, the model applies to interactions between Artificial Agents (AAs), which allows the decision calculus to be fully rational.

I summarize several interesting features of the author's characterization of e-trust:

1. E-trust is *rational*, specifically appealing to Kant's regulative ideal of a rational agent, in which the agent chooses the best option for itself given a specific scenario and goal-orientation.
2. From the above, e-trust is both goal-oriented and action specific. In other words, it is permissible to trust an AA at one task but distrust them at another task; e-trust is not a global property given to AAs.

3. E-trust is a second-order relation that affects first order relations characterizing actions. For example, if AAs $A$ and $B$ transact via the sale ($S$) of some good ($g$), then $S(A, B, g)$. E-trust, $T$, is a second-order relation over transactions like $S$, meaning $T(S(A, B, g))$ will affect the conditions under which $A$ sells $g$ to $B$.
4. E-trust has the property of minimizing the trustor's effort and commitment to the achievement of a given goal. This happens by delegation of an action to the trustee, together with limited supervision of the trustee. The less a trustor trusts, the more they will supervise, or even replace, the actions of the trustee.

Roughly, an algorithm for assessing trustworthiness between AAs is spelled out like the following: an AA calculates the ratio of successful actions to total actions performed by the potential trustee to achieve the same or similar goals [11] (7). The technical meaning of "similar goals" is left out of the paper. Under this algorithm, e-trust is not calculable *a priori*, since the trustor needs previous actions from the trustee in order to assess it.

Lastly, the author indicates that extending the work to more complex cases, such as those where human agents (HAs) are either trustors or trustees, would be more complex. These cases bring attitudinarian and psychological factors into play, where previously only economic factors (rational factors) were relevant.

# 3 A Proposed Theory for Trust in XAI

## 3.1 A Model Theoretic Approach

After discussing three theories of trust in some depth, it's time to step back. I want to get clear about the metaphysical properties of trust in each of these models. In Walker-Buechner-Tavani, trust is a sort of spatiotemporal "zone" that affects actions undertaken between trustors and trustees. Because trust relationships are frequently "symmetric and reversible," we often find trustees acting as trustors, and trustors as trustees, for the same types of actions. The Taddeo-Primiero model says something agreeable, which is that trust is a second-order relation characterizing first-order relations over actions. If we think of the second-order relation as characterizing the "zone," and the first-order relations as expressing which actions frequently occur between trustors and trustees within said zone, the parallel is clear. Lastly, Nguyen's model considers trust as an "attitude" held by a trustor about a trustee. Nguyen also says that "in almost all cases, the scope of trust in $X$ will be restricted to particular functions of $X$" [6] (21). Here, again, it seems like "attitude" expressing something relational between the trustor and particular "functions of" the trustee.

There is a pattern between these three theories worth pointing out, which is that all instances of trust discussed seem to involve minimally (1) a trustor, (2) a trustee, and (3) an action the trustor trusts the trustee to take. So we can say that trust is the three-place relation

$$T(A, B, X)$$

which says "$A$ trusts $B$ to do $X$." I think this relational treatment coheres at least basically with all of the theories discussed above. For cases of so-called "simple trust" where the action is irrelevant, and $A$ "simply trusts" $B$, we can quantify over available actions $X$, as in

$$\forall X.\ T(A, B, X).$$

An important consequence of this formulation is that we can do away with the idea of *trustworthiness* altogether. A trustee $B$ is "trustworthy" just in case they're trusted by all relevant trustors in all relevant actions, or:

$$\forall A.\ \forall X.\ T(A, B, X)$$

So, trustworthiness is not anything more mysterious than trust – it's just a quantification over trust relations. We can remove the inner $\forall X$ if we want to say "$B$ is trustworthy in particular function or action $X$."

Now, I'd like to take a step, and argue that we ought to consider trust relations *model-theoretically.* I'll explain what this means. To understand, recall that the choice to trust is *up to the decision of the trustor* – in our formulation, $A$. In other words, $A$ gets to decide whether $T(A, B, X)$ or not. More specifically, $A$ might be inclined to believe $T(A, B, X)$ under certain conditions at a certain time, and then $\neg T(A, B, X)$ under different conditions at a later time – or vice versa.

It sounds plausible that $A$'s decision to trust be based on beliefs $A$ holds about $B$'s ability to do $X$, or about $B$ in general. Say that $A$'s relevant beliefs are contained in the set $C$, such that some $c \in C$ might be interpreted as "my friend said $B$ helped them with $X$ once" or "$B$ is a doctor" or "I don't know... they trashed the house last time we went out of town." If $\mathcal{M}_C$ is a model under which $A$ believes all $c \in C$, then $\mathcal{M}_C \models T(A, B, X)$ says that $A$ trusts $B$ to do $X$ *subject to their beliefs* codified in $C$.

I think this model-theoretic view, where models codify the trustor's belief states, affords some flexibility missing from the other theories. It allows me to trust entities in certain contexts but not in others – for example, after learning of their failures at that which I had entrusted them. Under this view, if

$$\mathcal{M}_C \models T(A, B, X)$$

only for $B$ to later betray $A$'s trust, we don't have to *nullify* our earlier assertion that

$$\mathcal{M}_C \models T(A, B, X).$$

We simply *also* say

$$\mathcal{M}_{C'} \models \neg T(A, B, X)$$

where

$$C' = C \cup \{\text{“}B\text{ betrayed me that earlier time”}\}$$

or something like that.

## 3.2 Coherence with Earlier Theories

I also think that the model-theoretic view of trust incorporates or subsumes the earlier theories in an acceptable way.

### 3.2.1 The Walker-Buechner-Tavani Theory

For the Walker-Buechner-Tavani theory, we can consider a "zone" of trust as bounded by a community of individuals and a set of relevant actions. $A$ and $B$ are drawn from the community of individuals, and $X$ from the relevant actions. If the zone is truly one of diffuse default trust, community members $A$ will tend to have mental models $\mathcal{M}_C$ such that

$$\mathcal{M}_C \models \forall B.\ \forall X.T(A, B, X).$$

Note that since the model is a mental model *internal* to $A$, this formulation allows for $A$ to be misguided in their trust. It also allows for

$$\mathcal{M}_{C'} \models \neg\forall B.\ \forall X.T(A, B, X),$$

say if $C'$ stands for some unusual belief set that would cause $A$ to distrust within their default zone under certain circumstances.

### 3.2.2 The Nguyen Theory

The model-theoretic formulation of $T(A, B, X)$ also coheres with Nguyen's "unquestioning attitude" account. If $A$ entertains some set of beliefs $C$ such that

$$\mathcal{M}_C \models T(A, B, X),$$

it makes sense to call such trust "unquestioning." The belief set $C$ is *sufficient* for $A$ to assess their trust in this particular instance, which is compatible with

$A$ not actively trying to add new beliefs into $C$ via questioning. On the other hand, if $A$ is actively questioning whether $B$ will $X$, given the set of beliefs $C$, it seems natural that $A$ does not trust but may be willing to trust after acquiring additional beliefs.

We can understand Nguyen's "questioning" as $A$'s attempts to append to or modify their beliefs in $C$. So, we might have

$$\mathcal{M}_C \models \neg T(A,B,X) \text{ and } \mathcal{M}_{C'} \models T(A,B,X),$$

which says that

1. $A$ didn't trust $B$ to $X$ given beliefs $C$,
2. $A$ questioned $B$'s ability to $X$,
3. $A$ received revised belief set $C'$ as a result of questioning, and
4. $A$ now trusts $B$ to $X$ given revised beliefs $C'$, and
5. $A$ no longer has incentive to question whether $B$ will $X$. If they did, they would have already done so before deciding upon the revised belief set $C'$ sufficient for trust.

### 3.2.3 The Taddeo-Primiero Theory

Finally, the Taddeo-Primiero view – that trust is a second order relation over first-order relations concerning actions – also coheres. This view is the most logically precise of the three theories we consider, so it might be the hardest to motivate as compatible. Nonetheless, Taddeo says the following in her 2010 paper on e-trust:

> If AAs $A$ and $B$ transact via the sale ($S$) of some good ($g$), then $S(A,B,g)$. E-trust, $T$, is a second-order relation over transactions like $S$, meaning $T(S(A,B,g))$ will affect the conditions under which $A$ sells $g$ to $B$. [11]

In order to relate $T(S(A,B,g))$ to $T(A,B,X)$, we can say that $S$ (the sale of some good) is equivalent to some collection of actions $X_S$, where each $X \in X_S$ is the discrete action of $A$ selling good $g$ to $B$ on some particular occasion. Thus,

$$T(S(A,B,g))$$

is equivalent to

$$\forall X \in X_S.\ T(A,B,X)$$

in our formulation.

One difficulty with equating the Taddeo-Primiero theory to our theory is that Taddeo-Primiero only concerns cases of trust between artificial agents. Critically, this means that the reasons for trusting are all entirely *rational*, which is not the case when we allow humans to take the part of trustors in our theory. For this reason the model-theoretic view of trust is more general than Taddeo and Primiero's theory, though I am still working out the details of how they relate or equate, given the logical similarities.

## 3.3 Connection to Explanation and Interpretability

Once we have established the model-theoretic view's coherence with the previous three theories discussed, we can turn to explainability and interpretability. Here I wish to argue that explainability and interpretability fit into our picture as operations on the contextual belief set $C$. In other words, the efficacy of model explainability (for trust) hinges on it's ability to contribute *difference-making beliefs* to an individual's decision to trust. This means that model explainability needs to be assessed *independently* for each relevant trustor, since the same explanation will affect different individuals differently.

As an illustration, take as an example some algorithm $B$ and a person $A$. $B$ executes some function $X$, and we would like $A$ to trust $B$ in doing $X$, or equivalently to trust the output of $B$'s doing $X$. Currently, $A$ has some set of beliefs about $B$ that we will call $C$, and let's suppose that $A$ does not trust in this scenario, meaning

$$\mathcal{M}_C \models \neg T(A, B, X).$$

We can understand an explanation, $E$, as informing $A$ of some details relevant to $B$'s doing $X$ – *$E$ explains how $B$ $X$'s.* In our formulation, this takes the form of revising $A$'s belief set $C$ to some new set $C_E$ – $A$'s beliefs after receiving the explanation $E$. The goal is for the explanation to be *difference-making*, which would mean precisely

$$\mathcal{M}_{C_E} \models T(A, B, X).$$

A similar argument can be made for transparency. Transparency is just some condition on the belief sets of potential trustors $A$, an idea which captures the requirement that transparency is specific to the *trustor* and not the trustee (in this case, an AI system). Transparency to some might be open-sourcing the software that trained a model, while this obviously does not count as "transparent" for someone uneducated in computer software.

# 4 Discussion / Conclusion / Future Work

In this section I will conclude the paper with some suggestions for future work, and applications for this model-theoretic view on trust that we've discussed. An obvious outstanding task is to characterize what makes a "difference-making" explanation or transparency guarantee. We've shown how explanation and transparency *can* fit into a model of trust that's sympathetic to other philosophical views, but have not dug into the particular *relationship* between explanation, transparency, and trust. This is a much bigger task.

# References

[1] Jeff Buechner and Herman Tavani. “Trust and multi-agent systems: Applying the ”diffuse, default model” of trust to experiments involving artificial agents.” In: *Ethics and Information Technology* 13 (Mar. 2010), pp. 39–51. DOI: 10.1007/s10676-010-9249-z.

[2] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey.” In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2018, pp. 0210–0215. DOI: 10.23919/MIPRO.2018.8400040.

[3] Richard Holton. “Deciding to Trust, Coming to Believe.” In: *Australasian Journal of Philosophy* 72.1 (1994), pp. 63–76. DOI: 10.1080/00048409412345881.

[4] Been Kim. “Interactive and interpretable machine learning models for human machine collaboration.” PhD thesis. Massachusetts Institute of Technology, 2015. URL: http://hdl.handle.net/1721.1/98680.

[5] Zachary Chase Lipton. “The Mythos of Model Interpretability.” In: *CoRR* abs/1606.03490 (2016). arXiv: 1606.03490. URL: http://arxiv.org/abs/1606.03490.

[6] C. Thi Nguyen. “Trust as an Unquestioning Attitude.” In: *Oxford Studies in Epistemology* (forthcoming).

[7] Giuseppe Primiero and Mariarosaria Taddeo. “A modal type theory for formalizing trusted communications.” In: *Journal of Applied Logic* 10.1 (2012). Special issue on Automated Specification and Verification of Web Systems, pp. 92–114. ISSN: 1570-8683. DOI: https://doi.org/10.1016/j.jal.2011.12.002. URL: https://www.sciencedirect.com/science/article/pii/S1570868311000668.

[8] Pearl Pu and Li Chen. “Trust Building with Explanation Interfaces.” In: *Proceedings of the 11th International Conference on Intelligent User Interfaces*. IUI ’06. Sydney, Australia: Association for Computing Machinery, 2006, pp. 93–100. ISBN: 1595932879. DOI: 10.1145/1111449.1111475. URL: https://doi.org/10.1145/1111449.1111475.

[9] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier.” In: *CoRR* abs/1602.04938 (2016). arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938.

[10] Peter Strawson. “Freedom and Resentment.” In: *Proceedings of the British Academy* 48 (1962), pp. 187–211. DOI: 10.1073/pnas.48.1.1.

[11] Mariarosaria Taddeo. “Modelling Trust in artificial agents, a first step toward the analysis of e-trust.” In: *Minds and Machines* 20.2 (2010), pp. 243–257. DOI: 10.1007/s11023-010-9201-3.

[12] Margaret Urban Walker. *Moral Repair: Reconstructing Moral Relations after Wrongdoing*. Cambridge University Press, 2006. DOI: 10.1017/CBO9780511618024.

[13] Wikipedia contributors. *Explainable artificial intelligence — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Explainable_artificial_intelligence&oldid=1056933872. [Online; accessed 30-November-2021]. 2021.