Thesis Freewrite and Initial Literature Review

Stanford University Dept. of Philosophy

Ryan Othniel Kearns

11 October 2021

Contents

1 Freewrite

In our conversation last Tuesday, October 5, we touched on a number of questions related to different aspects of the thesis. The first bundle of questions has to do with fixing the scope of our consideration of trust within Explainable AI, and the latter launches more specific critiques on existing theories of trust and their candidacy for explaining the phenomenon under discussion in Explainable AI.

First, at a higher level, we addressed the discrepancy between "being trusted" and "being trustworthy." There is the question of whether "trustworthiness" as a predicate even needs mentioning when it comes to the issue of trusting AI. Trustworthiness is attractive as it assigns a property to the object "up for trusting," which allows us to more concisely talk about trust in relation to some object – otherwise, we would need to spell out the individual instances of trusting that apply that make for trustworthiness. However, we also talked about trust at this "instance level" being a four-place predicate: A trusts B with (or to do) x in context c. Specifically within XAI, A, x, and c seem to admit so much variability that a discussion of the trustworthiness of B seems like it would have to engage with individual instances of trusting to be useful, since not a whole lot would generalize. For example, for B to be trustworthy, full stop, we'd need to cover cases where A is both domain expert and layperson, which are already very different. Thus, I think the primary phenomena under consideration is $Trusts(A, B, x, c)$ and not $Trustworthy(B)$. So when XAI abstracts say something like "explaining our model gives us reasons to trust it," variables A, x, and c are most often left implicit or contextual, but it seems like they persist both on the explaining and the trusting side of the equation. We'll need to get specific about the intended values for these variables, of course – is it a goal for XAI that A be unbounded, e.g. that any common person should be able to receive an explanation? Or are certain explanations only available for very highly specialized domain experts like radiologists or court judges?

Another high level discussion point we addressed was the tension between system complexity and accuracy. My initial reaction was that within conversations around AI Explainability, the relationship between complexity and accuracy seems like straightforward tension. This is not always true – in certain domains simple systems like decision trees can yield all the accuracy you might want, and applying deep neural networks to the problem is introducing complexity for no gain. Also, overfitting is a phenomenon affecting all domains of machine prediction tasks, and indicates a point where initial complexity actually hinders accuracy. There is something to the point, though, that complexity and accuracy are at odds. Doshi-Velez et al., in their 2017 paper "The Role of Explanation in Algorithmic Trust," cite one often-heard complaint about AI Explanation that it may degrade system performance by incentivizing algorithm designers to pursue simpler yet more explainable architectures. This complaint needs to be addressed by AI Explainability in general – I don't think it holds water, because not only are complexity and accuracy non-trivially related, but complexity and explainability itself are not directly at odds either. Understanding the interplay of these terms will be important for any sensible proposal I might come up with.

Now getting more into the weeds with particular trust theories, a related tension we discussed was that between explaining or interpreting and trusting. This is a particular hurdle for adopting Nguyen's view, which says trust is a disposition to not question the trustee's actions and motivations. It seems like explaining or interpreting is in fact doing the exact opposite – actively questioning the trustee's abilities and decision-making calculus. Does this mean we come to trust algorithms *less* when we explain them? This cannot be totally right. Nguyen actually has something to say on this in a more recent preprint I stumbled upon, having to do with the irreconcilable tension between trust and transparency (Nguyen 2021, "Transparency is Surveillance"). I don't think that this latter paper addresses the objection outright, because the transparency Nguyen considers seems mostly to do with intrusion into communities of experts, not inanimate decision-makers like machine learning algorithms. Nonetheless, it will be helpful to consider these two papers in concert. *Transparency* is yet another ill-defined word tossed about freely in Explainable AI. The Forbes article I cited for my prospectus actually says that transparency ensures trust, in total contradiction to Nguyen's argument (see Kearns 2). So, part of a proper treatment of trust will involve contextualizing not just explainability but transparency as well as a separate aim with complex relational quality.

Finally, we talked about the possibility of beginning with a more naive calculus, like Bhattacharya's model of trust as expectation of positive outcome. I think such models, drawn predominantly from economics, can be helpful because they aren't baked into interpersonal accounts the way philosophical theories are, which can make them easier to disentangle from their original intended usages. Taddeo's model of e-trust is another model that operationalizes expected value, though like Bhattacharya's, it is restricted to rational decision-makers as subjects (Taddeo 2010). There is something to this, though – it represents an ideal use case we can build upon if we want to consider less rational and more normatively loaded cases, like humans trusting robot collaborators.

2 Literature Review

2.1 C. Thi Nguyen, 2021, "Trust as an Unquestioning Attitude"

This paper advances a view on trust that differs from the more popular philosophical accounts. Specifically, Nguyen moves away from the idea that trusting requires assigning *complete agencial states* to the trusted. Instead, trust is defined as adopting an "*unquestioning attitude*", in which we suspend deliberation and allow things to functionally integrate with our own agency. Key here is the idea that non-agents, like parts of the body, a smartphone, or the ground, can count as artifacts one trusts.

More technically, 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.

This two-tiered approach mirrors Michael Bratman's account of intentions and resolutions, specifically the idea of "cognitive inertia" (see e.g. Bratman 1987, *Intention, Plans, and Practical Reason*). Note in particular that the unquestioning attitude doesn't mean one will never question, just that one is generally disposed not to question.

Fundamentally, Nguyen thinks that we trust in order to expand our agency by functionally integrating pieces 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. Nguyen's paper includes many references to modern technology, which we "rely on" in a very particular way, characteristic of this account of trust. Nguyen defines the term **technological gullibility** as the disposition to too readily and irrationally integrate new technologies into one's agency.

Personally, I think Nguyen's proposal is a compelling place to begin for analyzing the use of trust within the Explainable AI literature, since it seems to mesh naturally with intuitive notions of trust for machines and generally non-human agents. There is a clear tension, though, in the fact that questioning attitudes are exactly those we take when attempting to unbox or explain models of interest, and this seems to limit the extent of "unquestioning attitude" that is possible.

2.2 Annette Baier, 1986, "Trust and Antitrust"

Baier's 1986 work on trust is one of the most referenced works in this subject into the present day. In this paper, Annette Baier sets out to distinguish the different forms of trust that may exist and try to formulate a *moral basis* for trust – practically speaking, a test to check whether a given instance of trust is moral.

Baier includes a bit of historical context in this paper, noting what she calls a "silence" on the topic of trust within moral philosophy. This is because, in her view, philosophers have been overly concerned with the types of interpersonal relationships best characterized by moral *contract*, that is, equal-footing relationships between autonomous individuals, mostly men. In Baier's view, this fixation causes us to ignore most of the instances where trust is relevant, namely in relationships with imbalanced power structures. She looks particularly at trust in infant-parent relationships, as well as patriarchal husband-wife relationships.

In general, Baier thinks that trust is nearly always present, even in fleeting interpersonal interactions, and we only really notice it when it suddenly goes missing. Baier distinguishes *trust* from "*mere reliance*" by noting that when trust is lost, we are liable to feel *betrayal.* This is a point visited in depth in McLeod's treatment of the subject (McLeod 2015).

Because of this point about betrayal, Baier concludes that trust is something like reliance with an expectation of *goodwill* (also explored in McLeod 2015 with the term "will-based theories"). Baier sees trust predominantly as a **three-place predicate**, where A trusts B *with* valued item C, and also predominantly as a relation between two individuals (she points out that this is a limitation of her theory).

Concluding, Baier develops what she calls a "moral test for trust," which she sketches out in (self-admitted) rough terms. A trusting relationship is *moral* according to Baier if the reasons for assigning trust or trustworthiness survive being made explicit to the trustee / trustor respectively. In other words, an immoral trust relationship would be on perpetuating because of deception on behalf of the trustee, or threat of force by the trustor.

Baier's account is influential and worth understanding, given that the majority of contemporary philosophers working on trust will cite this paper at some stage in their literature review. The account is, however, notably brittle, as it struggles to adhere well even to groups of people as recipients of trust or the *trustworthy* predicate, let alone non-human agents or systems involving humans and non-human collaborators. This brittleness stems from "goodwill," an anthropocentric property, being crucial to Baier's moral trust test.

2.3 Buechner and Tavani, 2011, "Trust and multi-agent systems: Applying the 'diffuse, default model' of trust to experiments involving artificial agents"

Jeff Buechner and Herman T. Tavani are philosophers, and their point in this paper is twofold. In the first section the two defend Margaret Urban Walker's conceptions of zones of default and diffuse default trust, developed in Walker's book *Moral Repair*. In the next section, the authors 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. The authors focus their attention on SPIRE, Shared Plans Intention Reconciliation Experiments from Grosz et al. 2002.

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 (Walker 85).

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 (43).

Walker, herself, considers a variation of this idea called "*diffuse* default trust" (Walker 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 important. I think I will need to read Walker for a more thorough understanding, but this theory is decently well cited and a good candidate for Explainable AI owing to its flexibility.

2.4 Wachter et al., 2017, “Counterfactual Explanations without Opening the Black Box”

This paper comes from the Harvard Journal of Law and Technology, and concerns counterfactual explanations in the context of GDPR and algorithmic decision-making. In summary, the authors propose counterfactuals as appropriate “explanations” of algorithmic decisions that might help data subjects contest decisions made about them.

Counterfactual explanations don’t “open the Black Box” – that is, they don’t concern the inner logic or mechanism of machine learning classifiers – so they avoid the difficulties of explaining technical ML concepts to laypeople. Instead, counterfactual explanations provide a subset of inputs that would result in the desired classification. The optimal counterfactual explanation accesses the “nearest possible world” where the classification decision was desirable.

According to the authors, counterfactuals can help *data subjects* (their term for people affected by algorithmic decisions) in three ways:

1. *Inform* them; i.e. help them understand why a particular decision was made;
2. Assist with *contesting* algorithmic decisions, mostly by instructing which aspects of one’s data profile were relevant in the decision; and
3. Understand what could be changed to achieve the desired classification in the future (assuming the model and the environment stay mostly fixed)

Much of the paper’s section IV talks about GDPR, and how the “right to explanation” in GDPR isn’t actually a legal guarantee. Sections 13-15 of GDPR are mostly concerned with notifying the individual that data collection or algorithmic decision-making is taking place – importantly, *ex ante*, otherwise there is no right to not be subject to an algorithmic decision. Most of the language in GDPR surrounding explanation has to do with the “logic involved” with typical classifications, not individual classification instances. The only clause to mention explanation by name is Recital 71, and the recitals are intended as recommendations over strict laws. Further, these sections don’t provide an explicit connection between the right to an explanation and the right to contest an algorithmic decision, though the former would likely inform the latter. This means there is little to no pressure on the algorithm’s designers to provide helpful information were someone to contest an algorithmic decision. This can be prohibitive in terms of resources and time, because data profiles for individuals can be massive. Counterfactual would alleviate this issue by providing a small set of actionable inputs that would change the classification decision.

2.5 Doshi-Velez et al., 2017, “The Role of Explanation in Algorithmic Trust”

This paper summarizes findings from the Artificial Intelligence and Interpretability Working Group at the Berkman Klein Center for Internet & Society. Here, Finale Doshi-Velez and colleagues consider a legally-motivated view on explanation of AI systems, particularly due to the right to explanation language present in GDPR. Thus, the desiderata for when to make algorithmic explanations available becomes legal necessity instead of other technical, philosophical, or ethical virtues. Nonetheless, the paper is helpful in pointing out several common concerns with AI/ML explanation, like the tradeoff with system accuracy or the accidental exposure of trade secrets or attack vectors like adversarial examples (1).

The authors also identify strengths and weaknesses of algorithmic explanations relative to human ones. One strength is the exactitude and invariability of a machine explanation; another is the absence of social pressure. One major weakness, and reason for the computational complexity of explanations in ML, is the fact that any *ex-post* explanation for machine behavior must have its granularity set ahead of time. For example, we cannot query for the explanatory importance of a newly considered variable that the algorithm was not trained to identify semantically, whereas humans can consistently adapt their explanations to consider new variables or scenarios.

2.6 Taddeo, 2010, "Modelling Trust in Artificial Agents, A First Step Towards the Analysis of e-Trust"

This paper develops a model of what the author calls "e-trust," which is trust occurring in online or digital environments. 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:

- 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.
- 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.
- 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.
- 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 (7). The technical meanings of the "ratio" and "similar goals" are 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.

I think Taddeo's e-trust account meshes appropriately with Nguyen's view in "Trust as an Unquestioning Attitude," specifically the final bullet point that trusting and supervising can be explicitly traded off between two fully rational agents. There can probably be some work done in generalizing e-trust to digital environments involving humans by adopting concepts from Nguyen, Bhattacharya, Walker, Buechner and Tavani.

2.7 Hardwig, 1990, "The Role of Trust in Knowledge"

Hardwig's 1991 paper seeks to challenge a conventional notion in epistemology that "knowledge rests of evidence, not trust" (693). Instead, Hardwig asserts that as modern knowledge acquisition increasingly comes to rely on teamwork and cooperation, it is our *trust* in others, and not our independent evidence, that serves as the foundation for our knowledge.

This is an uncomfortable account because, according to Hardwig, trust is "blind" (693). Yet in many settings it is the only way to acquire knowledge, particularly in modern science. Hardwig references a paper measuring the lifespan of charm particles, which took 280 person-years of work to complete. No individual human would be capable of coming to the experiment's conclusions on their own, and so with the conventional epistemological account it would be difficult to say that anyone *knows* the lifespan of charm particles. But this is uncomfortable, as we want the charm particle experiment to count as useful science. So, according to Hardwig, we need to grant the *team itself* sufficient evidence to justify the conclusion about charm particle lifespan, even though no individual team member may possess this evidence.

Hardwig states that trust factors into the origins of someone's knowledge, as well as the "context of justification" i.e. justifying that knowledge (696). Digging a bit more into trust, Hardwig develops what he calls the **principle of testimony**:

If A has good reasons to believe that B has good reasons to believe p, then A has good reasons to believe p.

We may also replace "has good reasons to believe" here with "knows" for a stronger version of the principle.
As a way towards trust, Hardwig analyzes the problem of coming to a belief through testimony or testimonial evidence. Sometimes, so claims Hardwig, the best reasons for justifying a belief will be testimonial, as in the case with charm particles, simply because good reasons resting on direct, nontestimonial evidence would be impossible to obtain. Hardwig again refers to the "blindness" of this kind of knowledge: the reasons for justifying p (and A's belief that p) are reasons that A does not have (699).
At this point, Hardwig is faced with a decision among three possible accounts:

1. There can no longer be knowledge in a lot of scientific communities relying on cooperation;

2. One can know p without having access to some of the best evidence for p;

3. Some knowledge is *actually* known by teams and not individual people on those teams.

Hardwig decides to argue for a version of 2, which is that A can *know* p without direct access to the best evidence, which requires a modification to our account of rational belief (699).

Following this exploration into belief obtained via testimony, Hardwig moves to investigating the properties of A's relationship to B that might imply "good reasons" to believe B's account that p (700). These qualities of B are, in short:

1. **truthful**: B is being honest
2. **competent**: B knows what constitute good reasons to believe in this domain
3. **conscientious**: B has done their work carefully
4. **adequate epistemic self-assessment**: B must not be misled about the limits of their own knowledge in the subject matter pertaining to p

So, A makes an assessment of B's character, both **moral** (truthfulness) and **epistemic** (competence, conscientiousness, adequate self-assessment) character. So, A *must trust* B insofar as they are relying on B and expecting a responsive form of goodwill or virtue from B to cement that trust. In this way I might think that Hardwig's account, though largely concerning trust in an epistemic setting, aligns with Baier's family of will-based accounts of trust (discussed in Baier, 1986 and McLeod, 2015). If A doesn't trust B, then B's testimony about p will not give A good reasons to believe p.

2.8 C. Thi Nguyen, 2021, “Transparency is Surveillance”

Nguyen’s paper puts forth several ideas about how transparency can have ill effects on expert communities when imposed as a kind of bureaucratic surveillance mechanism. Broadly construed, “transparency” according to Nguyen “indicates any process in which some entity makes information about its own activities available, to be used in further decisions and actions” (4). He breaks his critique of transparency into three arguments:

1. The *deception* argument, which says that experts will change at least the justifications of actions when reporting them, leading to a discrepancy between action and justification. This argument is drawn from Onora O’Neill, who’s main point with the argument is to indicate that trust and transparency often explicitly counter each other.

2. The *epistemic intrusion* argument, which at base says that in-domain experts can have their work intruded upon by external requirements for transparency, limiting the quality, integrity, and autonomy of their work.

3. The *intimate reasons* argument, which points out that authentic, publicly-accessible terms for reasoning within certain communities are impossible when the subject matter is intimate and only well understood by the community involved.

Of the three, the epistemic intrusion argument receives the bulk of the paper’s attention. Nguyen draws important distinctions between several modalities of transparency:

- *Assessment* transparency, in which procedural details are unveiled for the purposes of assessment, contrasts with *re-use* transparency, the less normatively loaded form characterizing, say, the free sharing of open source software.

- *Output* transparency, in which only visibility of discrete outputs like lives saved is made available, contrasts with the more intrusive *deliberative* transparency that articulates each decision in a procedure like a life-saving surgery.

- *Expert-aimed* transparency, in which procedures are unveiled within expert communities, contrasts with *public-aimed* transparency, in which the inexpert public is the intended audience.

Nguyen focuses his attention on *public-aimed deliberative assessment* transparency, considered by him (and me!) to be the most potentially nefarious. For *epistemic intrusion*, the extent of the intrusion is tiered. At first, *deception* is possible when transparency is required but still autonomous. For example, a doctor might assign a diagnosis in-line with their expert judgement, then fabricate a reason for the diagnosis on a form designed to provide transparency.

However, when the expert cannot control information flow in this way, they may be *limited* in what they can actually do under a transparent regime. Next, external overseers or stakeholders might introduce *incentivized guidance*, where they indicate value in following particular procedures that an expert would not choose for themselves. Finally, such guidance can be *internalized*, meaning the expert has subsumed these external values as their own and operates with them alongside their expert judgement.

At the most intrusive, transparency can thus debilitate experts' ability to act in the proper ways, as they're responsive to demands and values from those who know less about their field. As O'Neill and Nguyen jointly point out, there is a critical and unavoidable tradeoff between trust and transparency in knowledge- or ability-dense undertakings. If we trust too much, we open the door to bias, malpractice, and corruption. However, if we enforce too much transparency, we can stifle the autonomy and value-calculus present among experts within complex domains, hurting their ability to do meaningful work.

Nguyen's critique of transparency seems limited to instances where it concerns the undertakings of experts in certain fields like the sciences or arts. I do not think it has been intended to apply out of the box to model transparency, where intrusions like incentivized guidance simply will not apply. However, the tension between transparency and trust is very interesting, especially as a lot of the Explainable AI literature blindly suggests aiming for both as virtuous ends of their undertakings without accounting for this tension.

2.9 Bhattacharya et al., 1998, "A Formal Model of Trust Based on Outcomes"

This paper comes from the management science / economics world, and proposes a formal theory of trust that bridges existing theories from psychology and economics. According to the authors, many existing psychological theories assume inherent trustfulness as a personality trait, which downplays the situation-specificity of trust. At the same time, economic theories are overly concerned with situation-specific constraints and don't consider the interpersonal relationship in transactions where trust is important. So, a composite mathematical theory must encode both aspects of the situation and the trust-giving and -receiving parties.

The authors of this paper purport to do this. For a sequential scenario, in which agent 1 acts before agent 2, agent 1's trust of agent 2 can be given as:

$$\begin{aligned} T_{1,2}|a_1^* &= Pr(\mu_1 > 0|a_1^*) \\ &= \sum_{x \in \gamma_1} Pr(\alpha_1 = x|a_1^*) \\ &= \sum_{x \in \gamma_1} \sum_{a_2 \in A_2} F_1(x_1; a_1^*, a_2) \cdot c_1(a_2|a_1^*) \end{aligned}$$

- μ_1 is the "goodness" of an outcome x_1, and when $\mu_1 > 0$ the outcome is favorable for agent 1
- α_1 is the function taking an action (a_1) to an outcome (x_1). In a special case it is determinate, but we assume in general that it can be somewhat random
- A_2 is the set of all actions that agent 2 may take after agent 1 has taken their decided action (a_1^*)
- F_1 is the function taking joint actions from agents 1 and 2 to outcomes for agent 1. F_2 is defined similarly
- $c_1(a_2|a_1^*)$ is agent 1's *conjecture* as to what action agent 2 will take

There are additional technical details I'm omitting, including the case where the actions are simultaneous, and what happens when actions are determinate and outcomes visible (*hint*: trust boils down just to conjectures c). But the main takeaway is that this model defines trust as the product $F(\cdot) \times c(\cdot)$ of the outcomes function and the conjecture function across different possible worlds, weighted by the probabilities of those worlds. Mentioning back to the first paragraph here, it is the authors' opinion that economists focus too much on the former term, while psychologists focus too much on the latter.

The authors also define trust in words:

> Trust is an expectancy of positive (or nonnegative) outcomes that one can receive based on the expected action of another party in an interaction characterized by uncertainty (462).

This theory is one of so-called “outcome” trust, because agents should be more trustworthy when the expectancy of a positive outcome for them is higher. Importantly, this theory does not apply to trust in single-action instances, like trusting your car to start in the morning. It is however a simpler attempt at operationalizing trust relationships that may be fruitful for trust in AI up to a point.

2.10 Doshi-Velez et al., 2017, “Accountability of AI Under the Law - The Role of Explanation”

This paper comes from a conference of scholars at the Berkman Klein Center Working Group on AI Interpretability at Harvard. They recognize that due to two challenges – (1) that AI routinely make common-sense mistakes in decision-making that humans would avoid, and (2) that humans decide objective functions for AI, introducing an additional dimension of error into the system – there is a need to discuss AI explainability as a legal mandate in more detail.

The authors define accountability as “the ability to determine whether a decision was made in accordance with procedural and substantive standards and to hold someone responsible if those standards are not met” (2). Those “procedural and substantive standards” allude to legal mandate, and the authors bring up a number of examples of legally required explanations for humans and corporations. As one central thesis, the authors suggest that we can start legal requirements for AI explanations by asking of them what we ask of people. Importantly, the authors recognize the caveat that AI must be required to provide the explanation *ex ante* – while humans can conjure post-hoc rationalizations that suffice for explanations under the law, AI cannot generate explanations (at least, neither local explanations nor counterfactual explanations) without knowing the human-interpretable components it will be expected to give before the fact. The authors also balance critiques that requiring explanations from AI might hinder innovation, specifically because storing data and building systems for explanation is costly, or expose trade secrets.

In terms of explanations, the authors describe a form of local explanations largely inspired by Ribeiro et al., 2016 and others, as well as counterfactual explanations drawn mostly from Wachter et al., 2017 (p. 4-5). They also indicate that the desire for explanations changes depending on (1) the impact / significance of the decision, (2) the ability to contest a decision and assign blame, (3) our reasons to think an error occurred with the decision (p. 6-7).

2.11 Coeckelbergh, 2012, “Can We Trust Robots?”

This paper has two basic contributions.

First, the author distinguishes two views on trust to frame the question. In the first, called the *contractarian-individualist* approach, individuals make rational choices to trust based on their social situation and expectations. In the latter, called the *phenomenological-social* approach, the community or social scenario is prior to the individual, so trust is not as much decided as it is latent in specific social interactions. In other words, trust is *created* in the former view, *presupposed* in the latter. The author supports the latter view, which seems to be the less popular one on the topic of trust in technology generally.

The second contribution concerns the question of whether “trust” is something appropriate to ascribe to robots. Theories of trust are predominantly interpersonal theories (see even McLeod 2015, the SEP article on trust), except when we talk about trusting artefacts, where trust is an expectation an instrument will function. In this latter case, we talk about “trust as reliance” (54). Robots are a special case, in that you can debate whether they’re better understood as “quasi-others” (57) and thus interpretable under the human-human theories of trust. Being a “quasi-other” is dependent on the robot’s ability to be social, use language related to trust, and operate as a free agent. In the event that robots don’t fulfill these criteria, Coeckelbergh suggests we resort to a “functionalist, performance criterion” (58): can we trust the robot to do what it’s expected to do / what we’ve intended it to do?

2.12 McLeod, 2015, "Trust" (Stanford Encyclopedia of Philosophy)

Being an SEP article, this piece is a summary / literature review of various philosophical arguments surrounding the central question, **"When is trust warranted?"** Here warranted is intended to include "justified," "well-grounded," and "plausible."

Exploration into this question takes us into three main domains, the first concerning the *nature* of trust (and trustworthiness) itself, the second concerning the *epistemology* of trust, and the third concerning trust's *value*.

First, there is philosophical debate concerning the nature of trust, dating back (sort of) to Baier's 1986 account, "Trust and Antitrust," which sets the first example of a **will-based** account of trust, stating essentially that trust is reliance plus an expectation of goodwill from the trustee. This is one of a family of **motives based** theories, which generally purport that trust differs from reliance and other attitudes by virtue of the trustee possessing the right sort of motivations. Another motives-based account is Russell Hardin's **encapsulated interests** theory, which says that trustworthy people are motivated by self-interest to maintain their trustworthiness in the eyes of others. In addition, there's a **virtue** account of trustworthiness that paints a picture of a sort of "general" trustworthiness, stating that a person is trustworthy by virtue, i.e. it is baked into their character.

Separate to motives-based theories are **risk-assessment** theories, which liken trust to an assessment of low risk in relying on the trustee, meaning the trustee is assumed willing to do what they're trusted to do. These theories do little to differentiate trust from *mere* reliance (as it's called frequently in the literature) and so are criticized.

Finally, we can theorize that the nature of trust is neither willingness to perform nor a certain motivation. These theories are typically based on a sort of **normative expectation**, wherein the trustor takes a certain *stance* towards the trustee, like Richard Holton's "participant stance." One prominent normative expectation theory is the **trust-responsive** theory, which states that being trustworthy involves having the right sort of response to being trusted to do X.

All of these theories have various drawbacks that are discussed in the article, so it's not abundantly clear that philosophers have decided on a most appropriate nature for trust. Many philosophers have accepted a pluralist view on trust for this reason.

Moving on to the epistemological question. According to McLeod, the central epistemological question at stake here is, "Ought I to trust or not?" This is especially important when trust has potentially damaging ramifications. As such, the question of the **rationality** of trust is particularly central, which raises an issue immediately. Trust and rational reflection seem basically opposed, because fully rationalizing the reasons for trusting someone could be thought equivalent to reducing one's trust down to mere reliance. The author says that trust "inherently involves risk" and this causes the conflict (13).

Sometimes, we need to trust without the ability for rational reflection, as in with emergency room doctors. Other times, we may just bake trust into our daily ongoings, to the extent that rationally reflecting upon *all of it* is impossible.

Two central debates in the epistemology of trust seem to be the **truth-directed** vs. **end-directed** accounts of rationality, and the **internalist** vs. **externalist** accounts of rationality. "End-directed" rationality is a sort of strategic rationality that serves non-epistemic aims, like putting the trustor at ease with the situation. "Internalist" rationality purports that the reasons for trusting need to be justifiable to the trustor themselves, whereas the "externalist" argument suggests that the rationale for trusting depends just on the "epistemic reliability" of its cause, and need not be known to the trustor (15).

This section also spends some time addressing the social and political climate often serving as a backdrop for trust, which we may want to affect the **default stance** people bring to trusting others.

Lastly, the value of trust can be intrinsic or instrumental, though most philosophers only care to identify the instrumental values. For one, trust can be helpful for **cooperation**, and even essential to cooperation if Friedrich and Southwood's theory, that trust is essential to **promising**, is true. Philosophers also identify meaningful relationships / attachments, knowledge (scientific, moral, or in general), and autonomy as instrumental values of trust.

I will be coming back to this summary a lot, I think. It made sense to read this SEP article earlier in the process, since it provides me with a deep reading list with which to continue the dive into trust philosophically, particularly concerning the epistemology and rationality of trust and the values of trust, especially cooperation in a team setting. I'm disappointed a bit that the dominant paradigm for philosophical work on trust is interpersonal trust, because I felt that Nguyen, Taddeo, Walker, and Buechner and Tavani's work on a more general form for trust was illuminating.