Ryan Kearns
Stanford Department of Philosophy
Undergraduate Honors Thesis Prospectus
Draft 1
2021 Oct 1

Explainable Artificial Intelligence (XAI) is the study and development of techniques to make black box AI systems and decisions more interpretable. Explanations of model behavior take a number of different algorithmic approaches (post-hoc counterfactuals[1], local-fidelity linear approximations[2], feature importance scoring, and so on) and target different levels of granularity (e.g., global systemic explanations or explanations of individual decision instances[3]).

Explanations for AI system decision-making are desired for multiple reasons: they provide *transparency*, groundwork for *accountability*, a starting point for algorithmic *recource*, and arguably bolster or at least assist in algorithmic *fairness*. Another virtue often touted in XAI research paper abstracts is that of algorithmic *trust*, which is the focus of this thesis. Most AI researchers seem content with the thesis that *explainable AI entails trustworthy AI*, without taking a hard look at

1. what trust *is*[4],
2. *what it is* we're trusting in the case of trustworthy AI, or
3. how explanation might entail trust, specifically
    a. whether the entailment between trust and explanation is direct or mediated by something like fairness, recourse, or robustness, and
    b. whether the entailment is necessary or merely correlative.

Stanford's Department of Computer Science recently taught a class entitled "Trustworthy Machine Learning."[5]. The following is taken directly from the course description on Stanford's "Explore Courses" website:

> This course will provide an introduction to state-of-the-art ML methods designed to make AI more trustworthy. The course focuses on four concepts: explanations, fairness, privacy, and robustness.

There was no unit or discussion of trust itself -- maybe the implication is that explanations, fairness, privacy, and robustness together guarantee trustworthiness. Maybe only a subset are required in certain scenarios. Whether either of those statements are true, I think it deserves to be spelled out.

Similarly, an influential paper in the XAI literature, its title beginning with "Why Should I Trust You?," offers the following important insight:

> 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*. (1)
> (emphasis in the original)

I think this is a great point, though nowhere in the paper is the word "trust" explicitly defined ("explain a prediction" and "interpretable," though, are both given rigorous technical definitions, on pages 1 and 2). This absence of a formal definition gives me the sense that we may be aiming at a moving target, or maybe not aiming at all.

As further motivation, here are some excerpts from recent paper abstracts in the XAI literature:

Doshi-Velez et al., "The Role of Explanation in Algorithmic Trust":

> When it comes to human decision-makers, we often want an explanation when someone makes a decision we do not understand or believe to be suboptimal. For example, was the conclusion accidental or intentional? Was it caused by incorrect information or faulty reasoning? **The answers to these questions permit us to weigh our trust in the decision-maker** and to assign blame in case of a dispute. (1)
> (emphasis mine)

Ribiero et al., 2016, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier'":

> Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, **quite important in assessing trust**, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. (1)
> (emphasis mine)

DARPA:

> Explainable AI—especially explainable machine learning—will be essential if future warfighters are to understand, **appropriately trust**, and effectively manage an emerging generation of artificially intelligent machine partners.[6]
> (emphasis mine)

Forbes:

> 53 percent believe that AI will always make decisions based on the biases of the person who created the instructions... This disconnect can be resolved with **transparency, which will in turn, ensure trust**.[7]
> (emphasis mine)

These quotations indicate various commitments from AI/ML practitioners as to what “trust” is and how we get there. Doshi-Velez, for example, seem to indicate that a proper explanation is all that is required for evaluating trust in the human-to-human case, ignoring some intuitive factors like the history of past actions and the power dynamic between the trustor and trustee. DARPA tags explainability as an "essential" (necessary) condition of trustworthy relationships, particularly for military applications of AI.

All told, it's not abundantly clear to me what artificial intelligence researchers and enthusiasts intend by this notion of trust, as they don't articulate how explainability is supposed to support or guarantee it. This makes sense if we consider that the import of this research is a calculus for AI system *explanation*, not trust. Yet "trusting" is itself a complex epistemological state, and "trustworthy" a property that I think obviously means a lot more than "explainable" or "transparent." I actually think it quite dangerous to infer that explainable AI systems are trustworthy ones in this way, particularly in sensitive settings like military applications, policing, criminal sentencing, loan and college application evaluation, and the myriad other ways black-box algorithms are present in the modern, digital world. I'm not accusing AI researchers of making this cognitive slip outright, just stressing that a lack of focus on trust in the literature makes the distinction blurrier than it should be. There's the conflation between the following two statements that I'm especially worried about:

(A) A good explanation of an AI system's decision-making *rationally entails trust* in the AI system's decision-making.

and

(B) A good explanation of an AI system's decision-making *yields a procedure for assessing trust* in the AI system's decision-making, subject to a particular definition of trust.

I have italicized the parts of each statement that differ. Statement (B), even if untrue, is what we should be aiming for; the mistaken assumption that (A) is substitutable I think is very dangerous. The difference is obvious in certain cases with human decision-makers. If you fix my cousin's car and I inquire "How?", and you explain that you tend to bang various engine components with a wrench until the ignition catches, I certainly wouldn't trust you with my own maintainence issues. On the other hand, a detailed explanation of my spinal surgery procedure is probably *not* the primary reason I trust the doctor undertaking it -- factors like her medical degree, prior experience, and the hospital's reputation seem more salient. In other words, statement (B) isn’t even always true. A search for *satisfactory conditions for trusting* AI systems seems to me as important and as complex as a search for satisfactory explanations of AI behavior, given that both cases differ from those involving human decision-makers in important ways. Such a search would also need to begin with an assertion of the *nature* of trust for our purposes.

Here is where I think some recent philosophical work can be beneficial. Serious study of trust in philosophical circles is actually quite new; the seminal paper on trust is Annette Baier's "Trust and Antitrust" from 1986[8]. Baier proposes that trust be construed as reliance plus an expectation of goodwill in the trustee's intentions, and also that trust is distinguished from "mere reliance" by the possibility of betrayal (254). Baier's view applies basically only to interpersonal settings, and does not generalize well even to *groups* of people we'd like to assign (non-) trustworthiness like co-authors, boards of directors, or government cabinets, let alone non-agencial decision-makers like machine learning algorithms. More recently, the philosophical literature has expanded to develop theories of trust amenable to non-interpersonal environments, notably digital environments[9]. Broadly speaking, and in a way I will articulate later, philosophers have gravitated toward conceptions of trust as a diffuse and background attitude underlying one's interactions with their environment, rather than an explicit contractual relation between two active participants[10][11][12]. These accounts, while not mainstream in the philosophical literature, relax Baier's and others' interpersonal constraints in a satisfactory way, allowing inanimate objects including technologies to be recipients of trust while still characterizing trust intuitively. Part of this thesis will include advocating *why* I think such "diffuse attitude" conceptions are appropriate models of trust for AI decision-makers.

Given the preamble above, a version of my central research question might be:

> What is the nature of trust in the XAI literature? What should it be? How can a critical conceptual analysis of trust in this setting, drawn from the relevant philosophical literature, promote more precise and aligned research objectives within the XAI community?

N.B.: the following section is speculative and may not be possible within the scope of a single paper on this subject. Nontheless, it's where my argument would continue, so I figure it worthy of inclusion at least on a draft at this early stage.

In the next section of the thesis, having established the nature of trust that we're aiming at, I will take a taxonomical approach through some of the predominant algorithmic techniques in explainable AI and measure them up to their goal of establishing algorithmic trust.[13] Such techniques tend to fall into a few major categories, including semantic feature saliency mapping and extraction[14], linear approximation[15], or post-hoc counterfactualizing[16].

---

[1] Wachter et al. 2017, "Counterfactual Explanations Without Opening the Black Box"
[2] Ribiero et al. 2016, "'Why Should I Trust You?' Explaning the Predictions of Any Classifier"
[3] Doshi-Velez et al., "The Role of Explanation in Algorithmic Trust"
[4] Lipton 2018, "The Mythos of Model Interpretability"

[5] See http://web.stanford.edu/class/cs329t/
[6] https://www.darpa.mil/program/explainable-artificial-intelligence
[7] https://www.forbes.com/sites/madhvimavadiya/2020/09/17/dont-blame-the-algorithm-trust-it/?sh=1ef8fc0e4830
[8] Baier 1986, "Trust and Antitrust"
[9] Taddeo 2010, "Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust"
[10] ibid.
[11] Buechner and Tavani, 2011, "Trust and multi-agent systems: Applying the "diffuse, default model" of trust to experiments involving artificial agents"
[12] Nguyen 2021, "Trust as an Unquestioning Attitude"
[13] N.B. by "algorithmic trust" I specifically mean a scenario with the black-box system *itself* as the recipient of trust. I ignore cases where the whole system (say of AI system plus human review board plus legal right to recourse) may be separately worthy of trust.
[14] Been Kim, Julie Shah, and Finale Doshi-Velez. Mind the gap: A generative approach to interpretable feature selection and extraction. In Advances in Neural Information Processing Systems, 2015b.
[15] Ribiero et al. 2016, "'Why Should I Trust You?' Explaning the Predictions of Any Classifier"
[16] Wachter et al. 2017, "Counterfactual Explanations Without Opening the Black Box"