Trust in Explainable Artificial Intelligence
with Application to Algorithmic Recourse

Ryan Othniel Kearns

19 November 2020

1 Introduction

Much of the literature on explainable AI (xAI) asserts that **explainability** somehow promotes, underwrites, or entails **trust** in AI systems (Doshi-Velez et al. 2017; Karimi et al. 2020; Hendricks, Hu, et al. 2018; Hendricks, Akata, et al. 2016; Ribeiro, Singh, and Guestrin 2016; Wachter, Mittelstadt, and Russell 2017). Work on algorithmic recourse, in particular, has made this claim (Wachter, Mittelstadt, and Russell 2017; Karimi et al. 2020). Yet to my knowledge, the connection between explanation and trust – that is, the necessary and sufficient conditions under which explanation might entail trust – has yet to be formally explored. This connection is nontrivial, because trust, itself, is nontrivial. Trust in algorithms, along with rigorous mechanisms for asserting it, will be instrumental in justifying the use of AI in military, medical, and civic settings, among others.

Moreover, explanation and trust seem to concern essentially different things. Explanation is fundamentally *ex post*, defined for specific outcomes after an event has occurred (Karimi et al. 2020; Wachter, Mittelstadt, and Russell 2017; Buesing et al. 2019; Bareinboim et al. 2020). Trust seems fundamentally *ex ante*, concerning forward-looking conjectures about events and actions (Bhattacharya, Devinney, and Pillutla 1998; Taddeo 2010; Coeckelbergh 2011). Thus, I believe the connection between explainability and trust should be met with skepticism *prima facie* and examined in more detail. To accomplish this task, we require clarity on the type of trust we're looking for in AI systems, and ultimately a formal treatment of said trust on par with that of explanation.

In this paper, I use a basic operationalization to show how algorithmic recourse fails to promote trust in a way that can be made technically rigorous. Since recourse relies fundamentally on counterfactual explanations, this result provides reason to doubt a straightforward relationship between explanation and trust, as has been claimed. I then advocate for a more formal treatment of trust in xAI in general, and survey potential starting points drawn from philosophy, psychology, and economics.

2 Algorithmic Recourse

Today, algorithms are employed to make a number of important decisions. Some of these decisions have negative consequences in people's lives, and this motivates the need for algorithmic recourse. Given a decision made on one's behalf by an algorithm, **algorithmic recourse** is described as "the systematic process of reversing unfavorable decisions by algorithms and bureaucracies across a range of counterfactual scenarios" (Venkatasubramanian and Alfano 2020, 284).

Wachter, Mittelstadt, and Russell 2017 were the first to employ counterfactual explanations for contesting algorithmic decisions. They define a **counterfactual explanation** as the following:

> Score p was returned because variables V had values $(v_1, v_2, \dots)$. If V instead had values $(v'_1, v'_2, \dots)$, and all other values had remained constant, score p' would have been returned (848).

Say an individual applies for a loan, and an algorithm rejects the application. Should the individual want to understand why their loan application was rejected, a counterfactual statement such as

> "Had your reported annual income been $60,000,
> you would have received the loan."

functions as an explanation.

If the individual wishes to contest the algorithm's decision, counterfactual information can be helpful in at least two ways. First, the reported information could be incorrect – that is, the individual's actual income may exceed $60,000. In this case, the decision was unfair because the provided information was inaccurate, and the individual can contest on these grounds. Here, the counterfactual saves the individual from the costly alternative of auditing their entire data profile for errors. Second, the individual may have an opportunity to apply for another loan in the future. In this case, the counterfactual allows the individual to *know what it would take* to receive the loan the next time around – they have to get their annual income above $60,000[1].

Karimi et al. 2020 extend Wachter, Mittelstadt, and Russell 2017's work by distinguishing between **explanations** and **recommendations** in algorithmic recourse. An explanation answers the question: "Why was a particular decision made?" Once again, the answer is given as a counterfactual explanation relying on contrast classes, like the one above (Lipton 1990). The authors define the **nearest contrastive explanation** as, roughly, the nearest possible world in which the classification decision is favorable.

A recommendation answers a harder question: "What could I have done differently to influence the decision the algorithm made?" The authors define the

[1] In practice, Wachter, Mittelstadt, and Russell 2017 note that multiple counterfactuals may be provided, so multiple paths to recourse may be possible.

minimal consequential recommendation as, roughly, the lowest cost course of action needed to reach a possible world where the classification decision is favorable. Thus, solving for the minimal consequential recommendation yields a free contrastive explanation, though not necessarily the nearest one. This should not matter, though, as the authors' intention is to minimize cost-to-recourse for the individual.

It will help to devise a simple technical formulation. An individual is represented by the feature vector $x \in X$. A classifier is a function $f : X \to Y$ where $Y = \{0, 1\}$ in the binary case[2]. Say, arbitrarily, that 1 is the favored outcome for the individual represented by x. In our loan example, x may include features like annual income and credit score, while $y = 1$ indicates that a loan was approved, and $y = 0$ indicates that a loan was rejected.

Given an input vector $x \in X$ and classification $y \in Y$, define the **recourse function** $r : X \times Y \to Y$, which yields[3] a recourse prediction $\hat{y}$ via a modification[4] $\hat{x}$ of x such that $\hat{y} = f(\hat{x})$. $\hat{x}$ is identical to the vector $(v'_1, v'_2, \dots)$ in Wachter, Mittelstadt, and Russell 2017's definition of a counterfactual explanation. The **recourse case**, as captured in Venkatasubramanian and Alfano 2020's definition, is when $f(x) = 0$ yet $r(x, f(x)) = 1$, meaning the original negative outcome has been undone.

3 A Survey of Trust

Our next goal is to operationalize trust in the algorithmic recourse setting, ideally in a way that respects other formalizations of trust in the literature.

Trust has many definitions across fields including economics, psychology, and philosophy. Bhattacharya, Devinney, and Pillutla 1998, in their attempt to unify economic and psychological perspectives, define trust as "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). In a similar vein, Mayer, Davis, and Schoorman 1995 define trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other party will perform a particular action important to the truster" (712). Taddeo 2010, in a paper on trust between agents in digital

[2] This treatment can be easily extended to multiclass settings, though I do not consider them here.

[3] In Karimi et al. 2020, recourse produces recommended actions, or interventions, suitable for reaching the modified input $\hat{x}$. For the scope of this paper, I simplify recourse by assuming the actions are implicit in the counterfactual recommendation, i.e. "Had your income been $60,000 annually, you would have received the loan" implies that one should take action to raise their annual income. This simplification blurs the distinction between explanation and recommendation that Karimi et al. 2020 make in their paper, though the specific technical details will not matter for this argument.

[4] When $f(x) = 1$, it's fine to assume $x = \hat{x}$. In this case, there is no "unfavorable decision" to reverse, and $r(x, 1) = 1$ for any $x \in \{x : f(x) = 1\}$.

contexts, says trust is "understood as a measure that indicates to the truster the probability of her gaining by the trustee's performances" (7). Finally, Coeckelbergh 2011, on a particular sort of instrumental attitude he calls "trust as reliance," says: "We sometimes say of an artifact that we trust it. What we mean then is that we expect the artifact to function, that is, to do what it is meant to do... to attain goals set by humans" (54).

A handful of commonalities emerge, particularly the use of words like "expect", "outcome", "gain", and "goal". These features can be intuitive starting points when devising a formal model. I highlight three I think are important:

1. Trust is **conjectural**. A formal calculus for trust should consider predictions about the trustee's behavior, relative to the truster's epistemic state.

2. Trust is **action- and context-specific**. This is just to say that trust between two agents *in general* is undefined – a truster trusts a trustee relative to a specific action at a specific occasion[5].

3. Trust is **outcome-oriented**. This means that more probable and more favorable outcomes, relative to the beliefs and desires of the truster, should increase trust, and vice versa.

4 A Simple Mathematical Model

Given the three conditions above, I propose the following basic model, which mostly agrees with a simplification of Bhattacharya, Devinney, and Pillutla 1998's model. Given an individual i, a system s, and a context x, the extent to which i trusts s in context x is given by:

$$Trusts(i, s, x) \triangleq P_i(s(x) > 0)$$

Thus, the extent that an individual trusts a system is simply that individual's probabilistic belief (P_i) that the outcome determined by that system will be positive for them. P_i is chosen rather than P to reflect that the probability distributions come from individual i's epistemic state and are not objective.

Several caveats are in order. First, this model of trust is a **rational** one, in which the individual uses all available information to maximize their expected gain. Specifically, I follow Taddeo 2010 in employing the "Kantian regulative ideal of a rational agent" (2). This is despite empirical evidence from psychology and behavioral economics that people do not trust rationally (Berg, Dickhaut,

[5] For a dissenting account that trust is in fact general, see Rotter 1971. Rotter calls trust a "generalized expectancy" held by a truster that a trustee's promises are reliable. This stance is popular among psychologists who view trustfulness and trustworthiness as stable personality traits, rather than features of particular interactions. Future work should contend with these theories.

and McCabe 1995; Jordan et al. 2016), and philosophical claims that trust is better understood as a social presupposition, rather than an individual choice (Coeckelbergh 2011). I consider the rational setting to be an acceptable starting point for trust in xAI, though a complete theory should consider irrational cases. Human psychology is an important part of this puzzle. Second, note that while trust is essentially about positive *outcomes*, I am using $s(x)$ in my formula, which more accurately describes the *actions* of a system s across contexts $x \in X$. Conflating the two is fine for our purposes here, but a distinction should be made whenever actions have nondeterministic outcomes. Bhattacharya, Devinney, and Pillutla 1998's complete model provides a calculus for such cases.

5 Explanation & Trust in Algorithmic Recourse

We are now in a position to formally assess trust in the algorithmic recourse setting.

First, the extent to which an individual i trusts the recourse process itself is given by quantity

$$Trusts(i, r, x) = P_i(r(x, f(x)) > 0),$$

where r is the recourse function defined in section 2. Likewise, the extent to which that same individual trusts the *classifier*, f, is given by

$$Trusts(i, f, x) = P_i(f(x) > 0).$$

We can show[6] that this decomposes into a sum of two terms, minus their product:

$$\begin{aligned} Trusts(i, r, x) = Trusts(i, f, x) + P_i(r(x, f(x)) = 1 | f(x) = 0) - \\ Trusts(i, f, x) \cdot P_i(r(x, f(x)) = 1 | f(x) = 0). \end{aligned}$$

In words, trust in the overall recourse system depends entirely on two components:

1. $Trusts(i, f, x)$: trust in the classifier itself
2. $P_i(r(x, f(x)) = 1 | f(x) = 0)$: i's probabilistic belief in the **recourse case**, where recourse reverses a negative classification

For recourse to be favorable, either classification is favorable, or classification is unfavorable and recourse reverses it. The key is the following: without information on $P_i(r(x, f(x)) = 1 | f(x) = 0)$, just knowing $Trusts(i, r, x)$ doesn't guarantee any minimum value for $Trusts(i, f, x)$. In other words, $Trusts(i, f, x) = 0$ is compatible with any value $0 \leq Trusts(i, r, x) \leq 1$. Also, note that computing $P_i(r(x, f(x)) = 1 | f(x) = 0)$ requires knowing $P_i(f(x) = 0)$, which would

[6] See Appendix.

already determine $Trusts(i, f, x)$[7]. Thus, *ex ante*, even complete trust in the recourse process doesn't warrant any trust whatsoever in the classifier itself.

Karimi et al. 2020 appears to at least mildly conflate these two notions of trust: "Carefully reviewing assumptions and exploring... robustness issues in more detail is necessary to build trust in the recourse system, and *in turn, in the algorithmic decision-making system*" (emphasis mine). It is true that this distinction is less significant for cases like our loan application example, where recourse is a generally accessible feature. In such cases, trust in recourse is suitable even if we can't trust the classifier itself. Yet in many other applications of AI, trusting the classifier is of critical importance. There is no meaningful recourse for an autopilot failure or stock market flash crash. In these cases it is critical that we trust the *algorithm*, rather than the explanatory apparatus we build around it.

6 Conclusion

A formal definition of trust is focusing, since it requires us to specify the particular system that acts as trustee. Karimi et al. 2020, in conflating a classifier and its recourse apparatus as a singular trustee, make the misleading statement that providing recourse for a classifier promotes trusting that classifier. Even under a simple definition, we see that trust in certain AI applications, like autopilots and stock traders, cannot come from recourse. This definition of trust is clearly inadequate, as discussed in section 4, but it provides a place to start.

This result promotes the idea that a formalized notion of trust could be helpful to xAI writ large. The discussion in section 3 recommends starting points in economic, philosophical, and psychological theories. Also, the limitations identified in section 4, in particular irrational cases of trusting (Berg, Dickhaut, and McCabe 1995; Jordan et al. 2016) and trust in nondeterministic environments (Bhattacharya, Devinney, and Pillutla 1998), will need to be addressed in more complete models. Determining the precise conditions for warranted trust in algorithms could have implications for policy, law, and business and military strategies.

[7] Since $P_i(f(x) = 0) = 1 - P_i(f(x) = 1) = 1 - Trusts(i, f, x)$.

7 Appendix

Beginning with the definition for trust in recourse (r), we have:

$$\begin{aligned}
Trusts(i,r,x) =& P_i(r(x,f(x)) > 0) \\
=& P_i(r(x,f(x)) = 1) \\
=& \sum_{y \in \{0,1\}} P_i(r(x,f(x)) = 1 | f(x) = y) \cdot P_i(f(x) = y) \\
=& P_i(r(x,f(x)) = 1 | f(x) = 1) \cdot P_i(f(x) = 1) \\
&+ P_i(r(x,f(x)) = 1 | f(x) = 0) \cdot P_i(f(x) = 0) \\
=& P_i(r(x,f(x)) = 1 | f(x) = 1) \cdot Trusts(i,f,x) \\
&+ P_i(r(x,f(x)) = 1 | f(x) = 0) \cdot (1 - P_i(f(x) = 1)) \\
=& P_i(r(x,f(x)) = 1 | f(x) = 1) \cdot Trusts(i,f,x) \\
&+ P_i(r(x,f(x)) = 1 | f(x) = 0) \cdot (1 - Trusts(i,f,x))
\end{aligned}$$

Note that $P_i(r(x,f(x)) = 1 | f(x) = 1) = 1$, since this is the trivial case where $\hat{x} = x$, so:

$$\begin{aligned}
=& Trusts(i,f,x) + P_i(r(x,f(x)) = 1 | f(x) = 0) \cdot (1 - Trusts(i,f,x)) \\
=& Trusts(i,f,x) + P_i(r(x,f(x)) = 1 | f(x) = 0) \\
&- P_i(r(x,f(x)) = 1 | f(x) = 0) \cdot Trusts(i,f,x)
\end{aligned}$$

So:

$$\begin{aligned}
Trusts(i,r,x) = & Trusts(i,f,x) + P_i(r(x,f(x)) = 1 | f(x) = 0) - \\
& Trusts(i,f,x) \cdot P_i(r(x,f(x)) = 1 | f(x) = 0).
\end{aligned}$$

References

Bareinboim, Elias et al. (2020). “On Pearl’s Heirarchy and the Foundations of Causal Inference”.

Berg, Joyce E., John Dickhaut, and K. McCabe (1995). “Trust, Reciprocity, and Social History”. In: *Games and Economic Behavior* 10, pp. 122–142.

Bhattacharya, R., Timothy M. Devinney, and M. Pillutla (1998). “A Formal Model of Trust Based on Outcomes”. In: *Academy of Management Review* 23, pp. 459–472.

Buesing, Lars et al. (2019). “Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search”. In: *ArXiv* abs/1811.06272.

Coeckelbergh, M. (2011). “Can we trust robots?” In: *Ethics and Information Technology* 14, pp. 53–60.

Doshi-Velez, Finale et al. (2017). “Accountability of AI Under the Law: The Role of Explanation”. In: *ArXiv* abs/1711.01134.

Hendricks, Lisa Anne, Zeynep Akata, et al. (2016). “Generating Visual Explanations”. In: *ECCV*.

Hendricks, Lisa Anne, Ronghang Hu, et al. (2018). “Grounding Visual Explanations”. In: *ECCV*.

Jordan, Jillian J. et al. (2016). “Uncalculating cooperation is used to signal trustworthiness”. In: *Proceedings of the National Academy of Sciences* 113, pp. 8658–8663.

Karimi, Amir-Hossein et al. (2020). “A survey of algorithmic recourse: definitions, formulations, solutions, and prospects”. In: *ArXiv* abs/2010.04050.

Lipton, Peter (1990). “Contrastive explanation”. In: *Royal Institute of Philosophy Supplement* 27, pp. 247–266.

Mayer, R., J. H. Davis, and F. Schoorman (1995). “An Integrative Model Of Organizational Trust”. In: *Academy of Management Review* 20, pp. 709–734.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Rotter, J. B. (1971). “Generalized expectancies for interpersonal trust.” In: *American Psychologist* 26, pp. 443–452.

Taddeo, M. (2010). “Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust”. In: *Minds and Machines* 20, pp. 1–19.

Venkatasubramanian, S. and Mark Alfano (2020). “The philosophical basis of algorithmic recourse”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

Wachter, S., Brent D. Mittelstadt, and Chris Russell (2017). “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR”. In: *European Economics: Microeconomics & Industrial Organization eJournal*.