The Empathy Gap: Why AI Can Forecast Behavior But Cannot Assess Trustworthiness

Jason R. D'Cruz[1], William Kidder[2] and Kush R. Varshney[3]

[1]*University at Albany, State University of New York, 1400 Washington Avenue, Albany, NY, 12222, USA*
[2]*Status Labs, 701 Tillery Street, Austin, TX, 78702, USA*
[3]*IBM Research – Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY, 10598, USA*

Abstract

In previous work we have sought to characterize "trustworthy AI" (Varshney 2022, Knowles et al. 2022). In this work, we examine the case of AI systems that appear to render verdicts about our (human) trustworthiness, and we inquire into the conditions under which we can trust AI systems to trust us appropriately. We argue that the inability to take on another's perspective (henceforth, "empathy deficit") can both explain and justify our distrust of AI in domains in which AI is tasked with forecasting the likelihood of human (un)trustworthiness. Examples include the use of AI to make forecasts for parole and bail eligibility, academic honesty, and creditworthiness. Humans have an interest in ensuring that judgments of our trustworthiness are based on some degree of empathic understanding of our reasons and unique circumstances. The inability of AI to adopt our subjective perspective calls into question our trust in AI systems's assessments of human trustworthiness.

Keywords

empathy, moral reasoning, trustworthiness, artificial intelligence

1. Corrective for an Empathy Deficit

In the *Guardian* advice column, "Your Problems, with Anna Tims" (2020), a reader going by her initials, GH, expresses exasperation with a persistently low credit score that prevents her from buying a new home, keeping her and her son in rented accommodation that is both expensive and unsuitable:

> In 2016, a default was registered on my credit file by my bank after I missed three payments of my student overdraft. During this time, I was in an abusive and controlling relationship. I had no money of my own. Shortly before my son was born in the same month, the default was registered.
>
> I was forced to move hundreds of miles away to a new city. I wasn't able to divert post. My only source of income was statutory maternity pay, which was

given straight to my partner. When I left him and began to rebuild my life, I was made aware of the default and immediately repaid it at £60 a month.

In her compassionate reply, Tims points out that a credit score is a "blunt instrument" that registers the default "without the circumstances". She describes GH's predicament as "heart rending".

Where the credit scoring algorithm registered only the missed payments, Tims was able to apprehend the reasons for the missed payments [3]. Granted, a more sophisticated machine learning system fed with more data may eventually discern that whether a person meets financial commitments when they are mired in an abusive relationship is not predictive of whether they will meet financial commitments after the abuse has ceased. But what the algorithm will not register is the fact that being bullied and abused is a morally *excusing condition* for failing to meet a financial commitment. That is, the algorithm will not register that it is inappropriate to adopt the "reactive attitudes" [4] of blame and moral approbation toward a person in such circumstances because the missed payments do not reveal a lack of conscientiousness on GH's part. It is therefore not fair to assess GH as untrustworthy because of them.

Most machine learning (ML) tasks are "anti-causal", inferring causes (labels) from effects (observations) [5]. A machine learning system may discern that, say, a Tottenham resident is more likely to repay loans when Spurs are winning than when they are losing. But the fact that Spurs are losing is *not* an excusing condition for missing a payment.

For the purposes of accurate prediction and efficient allocation of resources, it does not matter whether the explanation for GH's default is morally excusing or not. But from GH's perspective, a diminishment of her credit score in non-excusing circumstances is unassailable even if unwelcome, while a diminishment in her credit score in excusing circumstances is patently unjust. Had GH's letter been about Spurs losing rather than about domestic abuse, we can imagine Anna Tims responding with cool dismissal rather than solidarity and helpful advice.

There is no reason to think that the class of considerations that matter for judgments of trustworthiness by empathic human observers is coextensive with the class of features that will increase an ML system's accuracy in forecasting behavior. The task of forecasting and the task of assessing trustworthiness are fundamentally different, and accordingly require different capacities. Even if decision subjects are granted [6]'s proposal for the 'right to be an exception' to an ML system, the ML system itself cannot be the one that makes an empathic decision regarding the trustworthiness of the individuals that have excusing circumstances.

Anna Tims did something that is routine for humans, but that no extant ML system, no matter how sophisticated, has the capacity to do. She was able to *empathize* with GH's position, putting herself in GH's shoes and inviting her readers to do the same. This allowed her to understand why GH was not able to pay her bills, and to arrive at the judgment that the circumstances were morally excusing. Her empathic response allowed GH to be *recognized* as an individual with a unique perspective and set of experiences. More than mere advice, Tims could offer validation.

Even ML systems that ascend Pearl's ladder of causation to the level of counterfactual imagination do not thereby display a capacity for empathy [7]. Interventions and counterfactual imagination may allow systems to better understand why GH did not pay her bills, but do not allow them to reason about moral excuses.

We naturally interpret decisions that preclude us from opportunities because of our perceived riskiness as derogations of our character. We want judgments about whether we will be trusted (with a loan, with bail, with work-from-home accommodations, etc.) to be sensitive to whether we *deserve* to be trusted or not. Such judgments necessarily take into account reasons in context, features that are intuitively apprehended by other humans via empathy.

We can of course envision an improved algorithm fed with more data and optimized with considerations of social justice in mind. In an updated schema, victims of a documented history of domestic abuse could be afforded an amnesty from derogatory marks on a credit report. But notice that the judgment that the algorithm ought to be updated in this way must be made by an entity with normative competence. We maintain that, in the context of assessments of trustworthiness, such competence includes being able to take up the perspective of the person being assessed.

2. The Role of Empathy in Calibrating Trustworthiness

Knowing a person's track record is essential in gauging their trustworthiness. But a track record still needs to be interpreted, especially when departures from a pattern are not easily legible. This will typically require understanding and assessing the reasons behind the relevant actions. Empathy allows us to uncover reasons that justify attitudes of trust or distrust. It is therefore a tool to gather evidence of (un)trustworthiness that cannot be accessed by merely assessing a track record. Deprived of this tool we are at a marked disadvantage in calibrating trust. When others lack this tool, we have grounds for skepticism about their capacity to assess *our* trustworthiness.

The kind of evidence that is afforded by empathy, crucial to making judgments of (un)trustworthiness, is conspicuously inaccessible to purely data- and pattern-driven "system 1" AI systems [8]. Even knowledge- and reasoning-driven "system 2" AI systems and hybrid neurosymbolic AI systems are not at a point that can interpret moral reasons in context. (Moral sense reasoning is more difficult than common sense reasoning, which AI systems are still far from mastering.)

It is true that attitudes of trust and distrust are continuously updated with feedback about the trustworthiness of one's trustees, which is often a matter of whether they behave as you expect them to. But a trustee's (un)trustworthiness is not always manifest in their performance or non-performance of the actions we expect of them. There is an enormous amount of room for interpretation.

Consider: I trust you to pick me up at the airport one morning but you are nowhere to be found. I later learn that the reason you did not show up is that your partner needed to be rushed to the hospital after suffering an anaphylactic reaction. It would be impossible for you to pick me up without endangering her health, and there was no time to warn me that you cannot come. You have not acted as I was counting on you to act, but you have not shown yourself to be untrustworthy in any measure. Quite the contrary: I would not be able to trust you with anything important in the future were you to abandon your partner in a moment of acute need just to meet your commitment to pick me up at the airport. This mundane story illustrates an important lesson about trustworthiness. Whether a person is trustworthy is not fully legible

in the bare facts of their action, that is, in whether they do what you expect of them. Rather, determination of trustworthiness requires interpreting information about the context and their reasons.

Consider now the alternate possibility that you show up at the airport, but you *would have* forgotten were it not the case that a radio story that morning mentioned airports and your memory was jogged. In this case you have acted in the way that I expected of you, but in a very close possible world you would not have. Moreover, the explanation for the closeness of that possible world is that you took insufficient care to ensure that you would not forget. Were I to learn that you showed up in part because you got lucky, I would not be inclined to judge that you are trustworthy or to trust you more deeply in the future.

That trustors shoulder an inescapable interpretive burden is a feature of any plausible account of trust and trustworthiness. On Jones' account, a person is "three-place trustworthy" only if she is competent with respect to the domain of trust and "would take the fact that A is counting on her [...] as a compelling reason for acting if counted upon" [9, p. 70–71]. Note that a "compelling reason" need not be an overriding reason. Compelling reasons may be outweighed by other reasons that are even more compelling. As Jones puts it, "There is a necessary vagueness about what it is to take a reason to be compelling, and there can be disagreement over whether an agent has in fact done this" [9, p. 71]. There is always the risk that a person is unfairly judged untrustworthy as a result of unreasonable expectations. There is also the risk that they disguise their untrustworthiness behind specious claims that they face pressing countervailing reasons.

On Hawley's commitment account, trustworthiness requires a person to ensure that their commitments do not outstrip their actions [10]. As Hawley puts it, "This requires judiciousness in acquiring commitments as well as doggedness in fulfilling commitments already acquired, independent of others' expectations. Trustworthy people must sometimes disappoint up-front by refusing new commitments, rather than violate trust later on: this is the moral 'power of no'." But even the power of 'no' has its limits. A person can be blameless in failing to foresee circumstances in which meeting one commitment will mean falling short with respect to another, as in the case in which one's commitment to caring for one's partner in an emergency outweighs one's commitment to pick up a friend at the airport. A broken commitment does not impugn trustworthiness *ipso facto*.

Meeting your commitments and being responsive to the dependency of those who count on you are crucial and closely related aspects to being a trustworthy person. We remain neutral on which of the two is more fundamental. More important for our argument is that appraisals of a person as trustworthy or untrustworthy require appreciating the full range of competing reasons that trustees face. Trustors need to be able to fit their trustee's actions into an interpretable story of trustworthiness or untrustworthiness. This requires of trustors that they project themselves into the deliberative position of their trustees, something that only conscious beings with normative competence (for the proximate future at least, only humans) are able to do.

Granted, there are domains in which track record alone is sufficient evidence of an AI system's competence to perform the task entrusted to it, such as detecting a melanoma from a dermoscopic image or determining the best move in a game of chess. We focus on domains in which AI is tasked with a particularly delicate and morally freighted commission: making forecasts that humans will naturally interpret as assessments of their trustworthiness [1, 2]. Earning trust in one's ability to assess trustworthiness requires credible demonstration of the capacity to

slow down and take on the perspective of the target of the assessment. This is a capacity that cannot be fully compensated by computational virtuosity or predictive accuracy, but requires a highly advanced level of knowledge and reasoning. In these domains, AI's empathy deficit both explains and justifies distrust of AI.

3. The Dishonor and Self-Fulfillment of Distrust

From the perspective of the deployer, AI systems are useful tools for making more accurate forecasts, managing risk and allocating resources most efficiently. But the stakes are very different for the individual whose trustworthiness is being assessed.

Being categorized as "too risky" is inevitably experienced as a judgment of one's character. To be denied opportunities because one is deemed "a bad bet" is deeply dishonoring. Distrust without warrant risks insulting, demoralizing, and disempowering, planting the seeds of behavior and character revision that does warrant distrust. Particularly in settings in which trusting behavior is the norm and distrusting behavior is therefore conspicuous, the risk of marginalization and dishonor represents significant, even dramatic, losses for the party who is wrongly distrusted and for the relationship as well. Being cut off from opportunity in a setting in which others are given the benefit of the doubt can open deep wounds that are slow to heal [11].

That we care very much about how we are perceived is an enduring feature of human nature. Adam Smith, the founding father of economics, thought that the desire for standing in the eyes of others was one of the most basic of human inclinations.

> Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favorable, and pain in their unfavorable regard. She rendered their approbation most flattering and most agreeable to him for its own sake; and their disapprobation most mortifying and most offensive.

It is easy to underestimate the degree to which a person who is the object of distrust is aware that they are so regarded. Slepian and Ames explore the effects of such awareness empirically [12]. Individuals have an intuitive sense of whether or not others perceive them as trustworthy, and tend to behave in a way that is consistent with how they imagine themselves to be perceived. They find that face-based judgments of strangers predict trustworthy behavior in a game involving trust and deception. But this link is mediated by subjects' expectations of how other people perceive them and in particular their expectation of whether they would be trusted by others. As they put it, subjects "seemed to have an awareness of how people would judge them, and they internalized these expectations and behaved in accordance with them" [12, p. 287]. This work extends previous work in social psychology on "self-fulfilling prophecies" [13] that focus on individual contexts (e.g. whether a student will conform to a teacher's high expectations). The meta-perceptions that Slepian and Ames find are not derived from a particular interaction but rather likely "from a range of contexts across a lifetime of treatment" [12, p. 287].

Just as trustworthiness is cultivated and reinforced by trust, so also untrustworthiness is cultivated and reinforced by distrust. A lamentable consequence of this recursive pattern of

trust- and distrust-responsiveness is that knowledge that a person is widely distrusted, whether or not such distrust is merited, provides (pro tanto) evidence that they are untrustworthy. Conversely, a distrusted person who has a grasp of the interpretive biasing that is a signature of distrustful attitudes has (pro tanto) reason to believe that entirely innocent actions are likely to be interpreted as indicating untrustworthiness. This person will not trust others to trust him.

The mere anticipation of such mis-recognition diminishes the motivation to be responsive to trust. All of this is the perfect recipe for a self-reinforcing and pernicious equilibrium. Doubtless there is a logic to distrusting those who are broadly distrusted, just as there is a logic to giving up on winning the trust of one's distrustful fellows.

We have moral reasons to disrupt the aforementioned pernicious equilibrium because it darkens our prospects for living in harmony, equality, solidarity, and moral community.

4. Empathy's Liabilities and Distinctive Value

One implication of the view we advance is that we may be able to cultivate greater trust in AI by improving a system's capacity to *display* empathy (for example, by creating chatbots that appear to share our emotions) [14]. However, in systems that assess trustworthiness, this deceptive strategy has serious practical, moral, and political liabilities. Empathy itself, and not just its appearance, provides grounds for trusting others' assessments of our trustworthiness. Cultivating trust in the absence of its grounds is deceptive and risky. Liao and Sundar describe how the trustworthiness of a technology can be communicated through "trustworthiness cues" that are embedded in interface features [15]. Misleading cues about an AI's underlying lack of empathic capacity risks overtrust.

But empathy's critics have argued that even empathy itself is risky. As Bloom puts the point, empathy's "spotlight nature renders it innumerate and myopic: It doesn't resonate properly to the effects of our actions on groups of people, and it is insensitive to statistical data and estimated costs and benefits" [16, p. 31]. Empathy is also susceptible to intergroup biases: We tend to be more capable of empathizing with those who share our group identities and less capable of empathizing with those who do not [17, 18].

Furthermore, nothing we say here rules out the possibility that algorithmic approaches to prediction will often fare better than human judgment in forecasting untrustworthy behavior. Indeed, [19] finds that a simple model that uses just two inputs — age, and past court dates missed — fares better than almost all human bail judges in predicting flight risk.

Yet, despite empathy's liabilities and the forecasting capabilities of algorithms, empathy's distinctive value justifies the inclusion of an empathic agent in the loop of decision making when it comes to judgments of trustworthiness. The distinctive value of empathy is its ability to recognize morally excusing circumstances such as those involved in GH's case. This value is directly tied to the "spotlight nature" of empathy that troubles Bloom. It is true that in many cases of moral judgment we do not want such a spotlight to guide our decisions, for example when determining the most effective way to distribute disaster relief. But when the question at hand is one of assessing the trustworthiness of *an individual*, a spotlight is what is needed. We want the nuances of an individual case to be considered. We want morally excusing circumstances to be recognized. Empathy is ideally suited to this task precisely because of its

spotlight nature, its ability to allow us to project ourselves into an individual perspective and appreciate the moral significance of unique circumstances.

Yet, given that ML systems will often outperform human beings when it comes to accurate forecasting in these domains, there is a tradeoff in requiring the involvement of empathy in assessments of trustworthiness. We acknowledge empathy's susceptibility to bias and do not dispute AI's forecasting potential, but we maintain that empathy is uniquely suited to enable the recognition of morally excusing conditions in a way that is inaccessible to current AI systems. Insofar as the capacity for this recognition is involved in a just assessment of our trustworthiness, then empathy will be involved in that assessment, and AI, no matter how accurate its forecasts, is currently unable to make the same sort of assessment.

Thus, the questions of forecasting accuracy and assessment of trustworthiness, while often entangled, can be distinguished. Just assessments of trustworthiness should involve more than forecasting based on track record. They should involve reasoning in a way such that morally excusing conditions can be recognized, and empathy enables this sort of recognition.

Of course, a "system 1" machine learning system equipped with enough facts could perhaps become better at identifying morally excusing conditions, but only if doing so led to better forecasts. It might turn out that those who experience abuse are more likely to meet financial commitments than similar individuals who have not experienced abuse. Or it might turn out that they are not. If the latter is the case, then the system will not consider the abuse as relevant to the forecast. But being dominated by an abuser is an excusing condition for failing to meet a financial commitment; it is a context with moral weight that requires normative competence to grasp. The problem with extant AI systems, even neuro-symbolic ones, is that they lack a means of determining conditions as *morally* salient. On the other hand, empathy allows us to appreciate moral salience by taking on an individual's perspective and contextualizing their actions.

In cases in which the question is precisely about the nuances of an individual's situation and perspective, merely gaining more facts about our behavior is not enough for normative competence when assessing trustworthiness. We want to be understood as more than a collection of these facts when the question is our trustworthiness, our character. In these cases, a deeper understanding of the reasons underlying our behavior is required, and empathy furnishes this deeper understanding.

5. Conclusion

In this paper we distinguish the task of forecasting behavior from the task of assessing human trustworthiness. The former is particularly well-suited to machine learning systems while the latter requires capacities that such systems lack. Machine learning systems make predictions about human actions in the same way that they make predictions about natural phenomena such as floods, earthquakes, and the weather. At times we take up this external perspective on ourselves, too, asking, "How am I likely to behave?" rather than, "What should I do?"

In contrast, when we assess an individual's trustworthiness, we assess not only their actions but also their reasons. This includes considerations of whether it is fair to count past lapses against them or whether the presence of excusing conditions means that they are "owed the

benefit of the doubt". Such evaluations require an understanding of *why* they acted in the way that they did. Empathy furnishes us with the ability to occupy the perspective of another human and understand their reasons. In empathizing we can ask questions such as: Is it reasonable to expect of this person that they act differently than they did? Is this action an expression of their moral character or merely a product of the circumstance? What was it that prevented them from acting conscientiously and what changes in themselves or in the environment might make a difference?

Machine learning systems that are fed with vast troves of data can excel at the task of forecasting behavior and will likely improve. But forecasting behavior is not sufficient for assessing human trustworthiness, which has an essential normative component. Such assessments require understanding a person's reasons given the particular circumstances they face. This we learn through empathy. Even the most advanced neuro-symbolic AI systems of today that combine ML with reasoning ability face an empathy deficit that will be difficult to overcome. Empathy presents a challenging frontier for AI research.

None of this implies that there is not a useful role to AI to play in cultivating and supporting human trustworthiness. Machine learning systems can be deployed to make forecasts about the efficacy of interventions aimed at helping people to do what they are trusted to do (for example, the efficacy of giving taxi vouchers to people dependent on unreliable public transportation or text alerts about upcoming court dates for people living in chaotic environments). Such a reorientation would have us abandon the idea that AI systems should serve as *arbiters* of human trustworthiness and to think of AI systems as tools to assist humans in cultivating and supporting social inclusion. Responsiveness to trust is an "ecological capacity" that depends both on context and on character [20]. AI is a powerful tool for helping us to create the conditions for trustworthiness.

Acknowledgments

References

[1] K. R. Varshney, Trustworthy Machine Learning, Independently Published, Chappaqua, NY, USA, 2022.

[2] B. Knowles, J. D'Cruz, J. T. Richards, K. R. Varshney, Humble machines: Attending to the underappreciated costs of misplaced distrust, in: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Arlington, VA, USA, 2022.

[3] T. McCoy, E. Pavlick, T. Linzen, Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 3428–3448.

[4] P. F. Strawson, Freedom and Resentment and Other Essays, Routledge, London, UK, 2008.

[5] N. Kilbertus, G. Parascandolo, B. Schölkopf, Generalization in anti-causal learning, arXiv:1812.00524, 2018.

[6] S. H. Cen, The Right to be an Exception in Data-Driven Decision-Making, Technical Report, Massachusetts Institute of Technology, 2022.

[7] J. Pearl, The seven tools of causal inference, with reflections on machine learning, Communications of the ACM 62 (2019) 54–60.

[8] G. Booch, F. Fabiano, L. Horesh, K. Kate, J. Lenchner, N. Linck, A. Loreggia, K. Murgesan, N. Mattei, F. Rossi, B. Srivastava, Thinking fast and slow in AI, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 15042–15046.

[9] K. Jones, Trustworthiness, Ethics 123 (2012) 61–85.

[10] K. Hawley, How to Be Trustworthy, Oxford University Press, New York, NY, USA, 2019.

[11] J. D'Cruz, Humble trust, Philosophical Studies 176 (2019) 933–953.

[12] M. L. Slepian, D. R. Ames, Internalized impressions: The link between apparent facial trustworthiness and deceptive behavior is mediated by targets' expectations of how they will be judged, Psychological Science 27 (2016) 282–288.

[13] R. Rosenthal, Interpersonal expectancy effects: A 30-year perspective, Current Directions in Psychological Science 3 (1994) 176–179.

[14] J. Howick, J. Morley, L. Floridi, An empathy imitation game: Empathy Turing test for care- and chat-bots, Minds and Machines 31 (2021) 457–461.

[15] Q. V. Liao, S. S. Sundar, Designing for responsible trust in AI systems: A communication perspective, in: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, Seoul, Korea, 2022, pp. 1257–1268.

[16] P. Bloom, Against Empathy: The Case for Rational Compassion, HarperCollins Publishers, New York, NY, USA, 2016.

[17] M. Cikara, E. G. Bruneau, R. R. Saxe, Us and them: Intergroup failures of empathy, Current Directions in Psychological Science 20 (2011) 149–153.

[18] M. Cikara, E. Bruneau, J. J. Van Bavel, R. Saxe, Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses, Journal of Experimental Social Psychology 55 (2014) 110–125.

[19] J. Jung, C. Concannon, R. Shroff, S. Goel, D. G. Goldstein, Simple rules for complex decisions, arXiv:1702.04690, 2017.

[20] V. McGeer, P. Pettit, The empowering theory of trust, in: P. Faulkner, T. Simpson (Eds.), The Philosophy of Trust, Oxford University Press, Oxford, UK, 2017, pp. 14–34.