Humble AI

Bran Knowles, Jason D'Cruz, John T. Richards, Kush R. Varshney

KEY INSIGHTS

(1) Decisions based on AI-generated predictions of a person's future behavior are naturally interpreted by decision subjects as assessments of their trustworthiness.
(2) Minimizing risk by denying opportunities to seemingly untrustworthy individuals increases the risk of trustworthy individuals receiving negative decisions.
(3) Experiences of being wrongly distrusted by AI contribute to public distrust of AI.
(4) Humble AI calls for AI developers and deployers to appreciate and mitigate these harmful effects.
(5) This approach will help to align the application of AI with human values and promote public trust.

1 INTRODUCTION

One of the central uses of AI is to make predictions. The ability to learn statistical relationships within enormous datasets enables AI, given a set of current conditions or features, to predict future outcomes, often with exceptional accuracy. Increasingly AI is being used to make predictions about individual human behavior in the form of risk assessments. Algorithms are used to estimate the likelihood that an individual will fully repay a loan, appear at a bail hearing, or safeguard children. These predictions are used to guide decisions about whether vital opportunities (to access credit, to await trial at home rather than incarcerated, or to retain custody) are extended or withdrawn.

An adverse decision—say a denial of credit, based on a prediction of probable loan default—has negative consequences for the decision subject, both in the near term and into the quite distant future (see the sidebar on credit scoring for an example). In an ideal world, such decisions would be made on the basis of a person's individual character, on their trustworthiness. But forecasting behavior is not tantamount to assessing trustworthiness. The latter task requires understanding reasons, motivations, circumstances, and the presence or absence of morally excusing conditions [14]. Although a behavioral prediction is not the same as an evaluation of moral character, it may well be experienced that way. Humans are highly sensitive to whether others perceive them as trustworthy [29]. A decision to withhold an opportunity on the basis that a person is "too risky" is naturally interpreted as a derogation of character. This can lead to insult, injury, demoralization, and marginalization.

Creators of AI systems can rather easily understand the costs of false positives, where people are incorrectly predicted to carry out the desired behavior. These costs include loan defaults that reduce profitability, criminal suspects at large who compromise public safety, and abusive caretakers who provide unsafe care to vulnerable parties. Seeking to avoid these costs—a stance elsewhere described as *precautionary* decision making—is consistent with a natural human tendency toward loss aversion, and can be the optimal strategy when the costs of false positives are much greater than the costs of false negatives [31]. That said, the costs of false negatives are often much harder to understand and thus fully appreciate. How much is lost, for example, by withholding credit from those who would actually repay their loans, by denying bail to those who would dutifully show up at a hearing, or by removing children from responsible caretakers? We argue in what follows that the consequences of false negatives are widely underestimated, contributing to cascading harms that animate widespread public distrust of AI. As a corrective we offer a notion of Humble AI. We propose that this humble stance can both align AI systems with moral aims of respect and inclusion as well as contribute to broader public trust in AI.

2 DISTRUSTFUL AI

In social science literature, trust is usually understood as a *willingness to be vulnerable* to the harm that would occur if the trusted party acts in a way that is untrustworthy, and distrust is conceived as a "retreat from vulnerability" [13]. AI systems do not have mental states like beliefs (about a person's trustworthiness), affective attitudes (fear, anxiety, etc.), or intentions (to approach or to withdraw). Nonetheless, these mental states, attitudes, and intentions may still be inferred by a human receiving a decision, not least because they can quite reasonably be ascribed to the humans who develop, own, and deploy AI systems. So while AI is not capable of thinking or feeling in ways humans do when they trust or distrust, those who create and deploy AI influence the system's tendency to accept or to avoid the risk of decision subjects acting in untrustworthy ways. For this reason, we find it apt to describe AI tuned to avoid the risk of untrustworthy behavior as characteristically *distrustful.* As we discuss below, this distrustful stance has consequences quite apart from whether decisions are statistically "fair".

2.1 Distrust contributes to misidentifying untrustworthiness

High decision accuracy can be characterized as some near optimal combination of true positives and true negatives. Again taking the example of granting a loan, the optimal case would have all granted loans being eventually repaid (true positives) with all rejected loans being properly classified as a future default (true negatives). From the perspective of the humans deploying this optimal system, it has correctly distinguished the trustworthy from the untrustworthy. In contrast, AIs inclined toward distrust are not just more likely to identify the untrustworthy, they are also more likely to *misidentify* people as untrustworthy. There are a couple of reasons why this is the case.

(1) *Amplification of weak signals.* When a machine learning classifier based on correlations rather than causal phenomena operates in a regime with high costs of false positives, the decision threshold gets pushed toward the tail of the likelihood functions where weak correlations have undue influence (see Figure 1). Here, decisions are more likely to be based on spurious correlations of the target variable with irrelevant features.

(2) *Decreased responsiveness to evidence of misrecognition.* In addition to these false negatives being more likely, they are also less likely to be detected *as* false. When focused on lowering the costs of false positives, AI creators are inclined to interpret the identification of a high proportion of "untrustworthy" individuals as validating their model rather than indicating a need for additional training or adjustment of the decision threshold.

Due to the above, distrustful AI is more likely to lead to denial of opportunities for individuals who *actually deserve* them, inflicting unnecessary harm in ways that can, quite understandably, increase the public's distrust of AI.

2.2 The inertia of distrust

It is a known phenomenon of interpersonal relationships that distrust tends to be self-reinforcing [13, 23], fueled by "distrust-philic" [19] emotional responses of both parties. Interestingly, here we see that AIs, entirely lacking in affect, are also implicated in what can be seen as distorted reasoning that reinforces untrustworthiness classifications. There are three components of this distortion.

(1) *Fundamental attribution error.* The reason credit score is such a widely used indicator is because it offers an easily legible and seemingly objective measure of a person's global trustworthiness. When an AI uses credit score as a feature in areas outside of credit worthiness, it is assuming that a) the score conveys relatively stable information about the person's disposition, and b) this information is useful in predicting behavior across a wide range of contexts. In social psychology, the tendency to over-emphasize dispositional over situational explanations of behavior is known as the *fundamental attribution error* [18]. When the output of one system is used as a feature for another, whatever situational information that may have influenced the first system is at least diluted if not eliminated in the second system. This leads to a systematic over-emphasis on disposition, which ultimately means that *a person who is miscategorized as untrustworthy by one AI has a greater chance of being miscategorized as untrustworthy by other AIs.*

(2) *Asymmetrical feedback.* When a person is trusted, they are given an opportunity to carry out the task they are trusted to do, and usually they are highly motivated to demonstrate their trustworthiness [4, 24, 27]. The resulting behavior generates new data that serves as "confirming" feedback to the trustor as they perform ongoing re-calibration of trust [32]. An individual's tendency to meet commitments or to fall short of them will, over time, influence AI's classifications of trustworthiness. In contrast, the distrusted lack the opportunity to become reclassified because they lack the opportunity to demonstrate how they would have responded were they to have been trusted. Distrustful strategies—retreating, withdrawing, avoiding reliance—lead to systematic under-trusting [15] by *reducing information about people's trust-responsiveness* that is needed to recalibrate misplaced distrust.

(3) *Reliance on proxies.* In contrast to a person who is distrusted, a trusted person not only receives a favorable decision, but that decision often creates opportunities that are characteristic of "trustworthy" individuals. Consider the lasting impacts of a decision regarding eligibility for rented accommodation: A trusted person is granted an apartment in an affluent neighborhood. A distrusted one is denied the same accommodation, so takes an apartment in a less affluent neighborhood, which may also have a higher crime rate. In a future decision about these two individuals, the one with a "better" postcode is more likely to be seen as more trustworthy [28] to the extent that proxy is used as a model feature in other AIs. Decisions based on proxies that are seen as being evidence of untrustworthiness perpetuate long-term disadvantage by *making it easier for the trusted and harder for the distrusted to be recognized as trustworthy.*

Here we see that, as in interpersonal relationships, distrust can create pernicious spirals that are very difficult to escape. Distrust by AI feeds itself insofar as it leads to system (by which we mean both within-system and system-of-system) feedbacks that *prevent re-trusting those classified as untrustworthy.* The difficulty a distrusted individual faces in establishing their trustworthiness makes miscategorization by AI more consequential than it may immediately seem.

2.3 People distrust those who distrust them

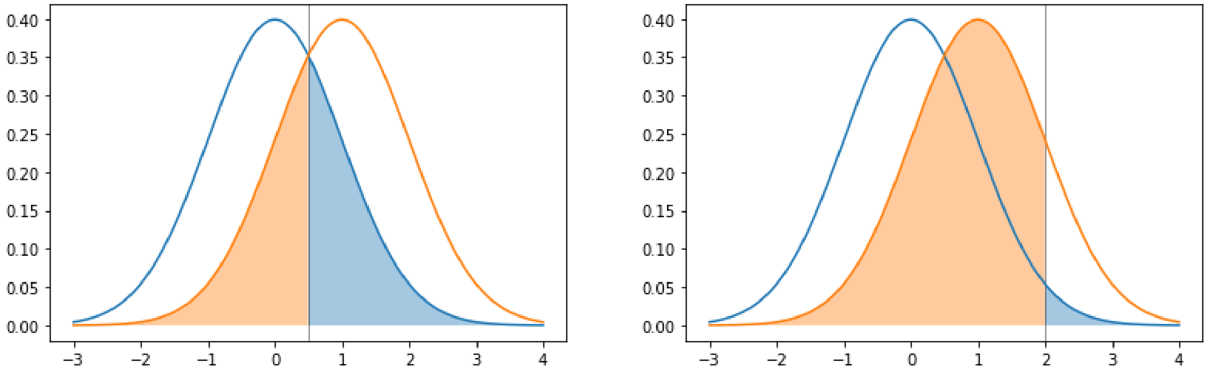
Having established that distrustful AI increases the likelihood of individuals being labeled as untrustworthy, let's now consider what it feels like to be on the receiving end of such labeling. Beyond the immediate consequences of an unfavorable decision, a decision that labels an individual as "too risky" (to receive a loan, to be granted bail, to retain custody) is a derogation of character which has important effects even when, or *particularly* when, it is unwarranted.

(1) *Resentment.* Frustratingly, trustworthiness does not *ipso facto* engender trust. A perfectly legitimate response to one's trustworthiness not being recognized is ill will toward the decision maker. When a person repeatedly experiences a mismatch between the evidence of trustworthiness required by AI and the evidence they are capable of providing, this naturally breeds resentment for being made to play a game whose rules are tilted against them. Resentment typically leads to "obsessively replay[ing]" one's injury, and also "cuts off search for possible mitigating factors or alternative explanations of it" [19], allowing distrust to build.

(2) *Demoralization.* It is valuable to have one's trustworthiness broadcast to other parties. As such, when a person is trusted, they have an incentive to respond to this trust so that this signal continues to be broadcast (this is known as the "trust-responsiveness mechanism" [27]). In contrast, a person who is distrusted lacks any such incentive because the signals they send are less likely to receive uptake. It is possible that being repeatedly identified as untrustworthy by AIs may demoralize a person in ways that spill over into how they comport themselves in their daily lives. In turn, this behavior may produce signals that serve as input to other AIs which are then more likely to identify untrustworthiness.

Just as trustworthiness is cultivated and reinforced by trust, so also untrustworthiness is cultivated and reinforced by distrust. The recursive pattern of trust- and distrust-responsiveness means that knowledge that a person is widely distrusted, whether or not such distrust is merited, is (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

SIDEBAR



(a)

(b)

Figure 1: (a) Equal costs for false positives and false negatives. (b) Higher cost for false positives than false negatives yielding a larger false negative rate and a smaller false positive rate. This pushes the decision threshold toward the right and yields many more people being classified as untrustworthy. Since the area of the orange region is now more than the area of the blue region, more people are being misclassified as untrustworthy. Also notice how the decision threshold is way to the right and at a point in the tail of the likelihood functions, which corresponds to weak and less informative data.

(pro tanto) evidence that innocent actions are likely to be misinterpreted. This person will not trust others to trust him. The mere anticipation of such misrecognition diminishes the motivation to be responsive to trust, generating a pernicious and self-reinforcing equilibrium [14].

(3) *Alienation and contempt.* When a person is distrusted for inscrutable reasons (as is the case with most AI systems), they may draw on other experiences of pain and injustice to construct what to them seems a plausible explanation [13]. Individuals who are unable to produce a sufficiently rich digital trail that matches expected patterns will find themselves unable to signal their trustworthiness and render it legible to AI systems. Those experiencing exclusion because they cannot satisfy AI systems may quite rightly feel contempt for the entire world represented by "AI"—the system of systems that circumscribe the new rules in society that appear to make it harder for certain individuals to succeed.

These mechanisms help explain how *distrustful AI provokes reciprocated distrust by the public.* As we see, there is an important affective dimension of trust and distrust—while one may reason about trust, emotions have a strong influence on that reasoning [19]. Contempt, for example, is a "totalizing emotion": "it focuses on the person as a whole rather than on some aspect of them" [19]. This is important when we consider people's broad distrust of AI to make *any* decision about them [8], as this reflects a collapsing of all AIs into a single threatening entity. These emotions may also reduce the individual's receptivity to the notion that AI *could* be trustworthy, e.g. even if improvements are implemented which correct for the original misrecognition. Thus we see that distrustful AI not only justifies distrust of AI, it also triggers affective feedback loops that intensify and entrench public distrust.

SIDEBAR: THE NATURE OF PUBLIC TRUST IN AI

Trust is typically assumed to be a singular and commonly understood phenomenon; however, any attempt to synthesize the vast literature relating to trust in AI quickly reveals trust as a high-dimensional concept [20]. Strategies for promoting trust in AI by decision makers, such as model explanations, do not necessarily translate to the trust needs of individuals subject to AI decision making, or those we refer to by "the public." We recognize we are, in fact, referring to a highly diverse category comprised of many different "publics" [16]; and we theorize that minoritized and marginalized groups may be particularly susceptible to negative effects of distrustful AIs to the extent that they contribute to multiplicative disadvantage.

Crucial to understanding how to foster public trust in AI is to recognize that their distrust is rooted in a very real asymmetry in the relationship: the AI is able to almost entirely minimize the vulnerability of the deploying organization while increasing the vulnerability of the decision subject. In this sense, the public's distrust could be understood as a response to the bureaucratic violence inflicted by hyper-efficient machines [6]. Careful adjustment of AI decision thresholds, and openness to a broader range of features signalling an individual's trustworthiness, can mitigate some of the consequences of misplaced distrust, but it will take more than this to restore the public's trust in AI. The public will need to see that those who develop and deploy AI systems genuinely respect the people who are affected by AI decisions, that they exhibit compassion overriding their desire for efficiency, and that they are committed to earning the public's trust.

3 A PROPOSAL TO EMULATE HUMBLE TRUST

To help address legitimate concerns about misrecognition by AI, we propose several affirmative measures inspired by the notion of "humble trust" [13]. Underlying humble trust is an awareness of and a concern for the harm caused by misrecognition. This means balancing the aim to not trust the untrustworthy with the aim to *avoid misrecognition of the trustworthy.* The practice of humble trust entails [13] (emphases added):

(1) "**skepticism** about the warrant of one's own felt attitudes of trust and distrust";
(2) "**curiosity** about who might be unexpectedly responsive to trust and in which contexts"; and
(3) "**commitment** to abjure and to avoid distrust of the trustworthy".

These principles have important implications for features, labels, costs, and thresholds of the decision functions in a machine learning system.

3.1 Skepticism: Confidence and verification

It is sensible to want to determine with the greatest possible accuracy who is trustworthy and who is not, but in doing so, one needs to be aware of the limitations of statistical reasoning and open to the possibility of getting it wrong. While machines are capable of finding usefully-predictive relationships within large volumes of data, we know that AI interpretation, like that of humans, is susceptible to uncertainty and failure. It is not always clear which feature or combination of features is most predictive of the desired behavior, nor how the available data relates to those features.

Key to avoiding overestimating the predictive capability of machines is *recognizing the information loss that occurs in selecting a set of features while ignoring others [11, Sec. 2.8] and the uncertainty that results [9].* What is the system not seeing by focusing on what it is focusing on? Are the data and model uncertainty too large? An appropriately skeptical stance would be to assume the model is missing information that could be relevant, to not be satisfied that a metric alone tells the whole story, and to actively seek out non-traditional evidence of trustworthiness as features that allow people to show themselves more fully. An active feature acquisition approach proposed by Bakker et al. operationalizes skepticism exactly along these lines and achieves fairness for both groups and individuals by continuing to seek additional features about individuals as long as the AI remains too uncertain [5].

Those who build and deploy AI systems must not lose track of the crucial distinction between predicting behavior and assessing trustworthiness. Being trustworthy is matter of responsiveness to being counted on and doggedness in meeting commitments. Such qualities are not necessarily derivable from a person's behaviors, which are influenced also by circumstance and opportunity. Being able to predict whether a person will meet a commitment does not imply an understanding of how easy or difficult it will be for them, or what mitigating or excusing conditions should be considered [14].

It is also worth noting that being forthright about such limitations does not necessarily, and should not, lead to reduced trust. In the wise words of Onara O'Neill, "Speaking truthfully does not damage trust, it creates a climate for trust" [26].

3.2 Curiosity: Trust-responsiveness

Part of being open to the possibility of having gotten a decision wrong is being curious about what might have happened if a different decision were made. This means creating opportunities for the AI to learn about the trust-responsiveness of people who fall below the decision threshold—in practice, extending trust to those who might betray it and seeing if the expectation of betrayal is fulfilled. Nuanced solutions to this problem (also known as the "selective labels problem") have been addressed in principled ways in the decision making and machine learning literatures. For example, Wei balances the costs of learning with future benefit through a partially-observed Markov decision process that shifts a classifier's decision threshold to more and more stringent positions as it sees more people that would normally have been classified as untrustworthy [32]. This may be seen as a way of conducting safe exploration wherein the system exhibits curiosity up to a point that does not induce undue harms [25]. It may also be seen as involving satisficing behavior, a decision strategy that aims for a satisfactory result but not necessarily the optimal one if curiosity were not a consideration [33].

3.3 Commitment: Investing in identifying and supporting trustworthiness

A common economic objective in deploying AI to make forecasts about human behavior is *lowering the costs of making a decision.* Eliminating humans from the decision process is tempting for this very reason. But AIs can do more harm then good when this sort of efficiency is pursued to the exclusion of other values, such as quality, fairness, and social inclusion. Commitment is exemplified by doing something *even when it is tempting not to*; so a commitment to avoiding distrust of the trustworthy means making adjustments to the model or wider decision making process even when those changes reduce the overall efficiency.

An embodiment of such commitment is the establishment of an institutional process (such as an AI Ethics Review Board) to carefully consider the costs of false negatives along with false positives, better aligning each AI system with core values of fairness and social inclusion. Another example of this commitment would be designing the AI to report when it is unsure and passing those decisions to humans who can be more deliberative, even if less efficient. Humans are able to "put themselves in the shoes of" decision subjects through empathy, identifying and evaluating mitigating and excusing conditions in a way that algorithms cannot [14].

SIDEBAR: THE EXAMPLE OF CREDIT SCORING

Credit scoring provides an example of feedback mechanisms that can fuel deeper or more widespread distrust. Credit scores are an indicator used for many purposes beyond determination of credit worthiness. A low credit score can make it harder for people to get a mobile phone, rent an apartment, or find affordable car insurance. Employers may run a credit check on candidates even for jobs that do not require the direct handling of money. Bad credit can block security clearances that affect military service members [12, 21] or other government workers. Some of these impacts can, in turn, feed back to further lower a credit score. This cycle can span generations, diminishing the prospects of offspring, as well as leading to entire communities being permanently labeled high-risk.

Credit scores are also frustratingly unstable—a person may have a lifetime of financial reliability, but if they abruptly begin to miss payments because of a setback beyond their control, say, because they are hospitalized with an illness or suffer some other unanticipated disruption in their life, their score may drop appreciably.

What might a commitment to humble trust suggest? One approach is being directly receptive to additional information by allowing an individual to provide notes and explanations. Credit rating agencies allow a Notice of Correction to be added to a credit report. It is limited, however, to 200 words. Ensuring it has been added to all credit agencies is tedious. Moreover, it may, in some cases, lead to more scrutiny of the adverse markers being noted in the correction. To our knowledge, there is also no automatic processing of such notices, so it is simply part of the credit report that can be viewed by entities assessing credit worthiness. Whether this is likely to heal a dysfunctional trust dynamic hinges on a Notice *being heard* and the evidence of that hearing being available to the individual who submitted the Notice. It also requires individuals to be both knowledgeable and proactive regarding their credit score, which is typically more challenging for exactly the same people whose life complications require closer attention to understand. But at least this opens the door to a conversation (in theory), and it represents an explicit acknowledgment that we cannot be satisfied that the score alone tells the whole story.

A second approach explicitly looks for features beyond those typically used in AI-based credit decisions. Examples include a system for unbanked customers in Kenya and other parts of East Africa basing decisions on mobile phone-based indicators [30]. A related approach has been adopted by commercial platforms such as Upstart.com [3] and FairPlay.ai [1], which have been successful in extending credit to the traditionally non-creditworthy without an increase in risk. They also tap a market segment that might be particularly responsive to being trusted—individuals who are highly motivated to demonstrate their trustworthiness to establish better credit. A final example is provided by Indigenous Business Australia [2], granting home and business loans based on culturally sensitive indicators of trustworthiness such as how an applicant has helped others in the community.

4 CONCLUSION

The allure of decision-making efficiency is powerful [22]. To an ever greater extent, a person's opportunities are circumscribed by AI-driven forecasts of their behavior. Consternation at this prospect is entirely reasonable. If we are serious about aligning AI with values of fairness and social inclusion, we must reorient our thinking about its appropriate use. AI is better suited to finding ways to cultivate and support trust-responsive behavior than to serving as an independent and objective arbiter of trustworthiness.

Humble trust does not imply trusting indiscriminately. Rather, it calls for developers and deployers of AI systems to be responsive to the effects of misplaced distrust and to manifest epistemic humility about the complex causes of human behavior. It further encourages them to look for (and provide opportunities for the future generation of) new signals of trustworthiness, thereby improving their ability to recognize the trustworthy. Finally, it suggests they look beyond the immediate efficiencies of decision making to consider the long term harms (both to individuals and AI-deploying institutions) of careless classifications.

For a business, misidentifing an individual as untrustworthy might mean losing a potentially profitable customer. Depending on the economic conditions, this may or may not impact an organization's near-term bottom line. The case is clearer from a moral perspective. Humble trust has an essential role to play in realizing justice and social inclusion. Democratic public institutions such as the criminal justice system and the social safety net cannot afford to compromise on such values.

While many applications of AI pose risks of exacerbating unjust distributions of trust, and therefore of opportunity, AI also offers unique mechanisms for resolving this very problem. It is possible to calibrate decisions made by AI systems with tools that are unavailable to the calibration of our own psychologies. Human attitudes of trust and distrust can be altered indirectly, but they are not under a person's direct voluntary control. This makes it difficult for human decision makers to adjust their personal attitudes of trust and distrust to align with moral aims. AI systems are different in this respect. The degree to which they are "willing to trust" can be directly manipulated by developers. The affirmative measures of the Humble AI approach promise to bring AI into better alignment with our moral aims so we may finally realize the vision of superior decision making through AI.

ACKNOWLEDGMENTS

This work is partially funded by the SUNY-IBM Research Alliance under grant number AI2102, the ESRC funded grant BIAS: Responsible AI for Labour Market Equality (ES/T012382/1), and by the Data Science Institute at Lancaster University.

REFERENCES

[1] [n. d.]. FairPlay - Fairness-As-A-Service. https://fairplay.ai. Accessed: 2022-12-31.
[2] [n. d.]. Indigenous Business Australia. https://iba.gov.au. Accessed: 2022-12-31.

[3] [n. d.]. Upstart Powered Loans: Personal, Car Refinance & Consolidation. https://upstart.com/, accessed 2022-12-31.

[4] Mark Alfano. 2016. Friendship and the Structure of Trust. (2016).

[5] Michiel A. Bakker, Duy Patrick Tu, Krishna P. Gummadi, Alex Sandy Pentland, Kush R. Varshney, and Adrian Weller. 2021. Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making using Confidence Thresholds. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 346–356.

[6] Veronica Barassi et al. 2021. David Graber, Bureaucratic Violence and the Critique of Surveillance Capitalism. Annals of the Fondazione Luigi Einaudi 55 (2021), 237–254.

[7] Peter L. Bartlett and Marten H. Wegkamp. 2008. Classification with a Reject Option using a Hinge Loss. Journal of Machine Learning Research 9 (Aug. 2008), 1823–1840.

[8] The Chartered Institute for IT BCS. 2020. The public don't trust computer algorithms to make decisions about them, survey finds. https://www.bcs.org/articles-opinion-and-research/the-public-dont-trust-computer-algorithms-to-make-decisions-about-them-survey-finds/.

[9] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 401–413.

[10] Yuheng Bu, Joshua K. Lee, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W. Wornell. 2021. Fair Selective Classification Via Sufficiency. In Proceedings of the International Conference on Machine Learning. 6076–6086.

[11] Thomas M. Cover and Joy A. Thomas. 2012. Elements of Information Theory. Wiley.

[12] Consolidated Credit. 2018. How Security Clearance Credit Check Rules Impact Many Military Service Members. https://www.consolidatedcredit.org/financial-news/security-clearance-credit-check/.

[13] Jason D'Cruz. 2019. Humble trust. Philosophical Studies 176, 4 (2019), 933–953.

[14] Jason R D'Cruz, William Kidder, and Kush R Varshney. 2022. The Empathy Gap: Why AI Can Forecast Behavior But Cannot Assess Trustworthiness. (2022).

[15] Detlef Fetchenhauer and David Dunning. 2010. Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. Psychological Science 21, 2 (2010), 189–193.

[16] Centre for Data Ethics and Innovation. 2022. Public attitudes to data and AI: Tracker survey. Available at: https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey.

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In Proceedings of the International Conference on Machine Learning. 1321–1330.

[18] Edward E Jones and Victor A Harris. 1967. The attribution of attitudes. Journal of experimental social psychology 3, 1 (1967), 1–24.

[19] Karen Jones. 2019. Trust, distrust, and affective looping. Philosophical studies 176, 4 (2019), 955–968.

[20] Bran Knowles, John T. Richards, and Frens Kroeger. 2022. The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency. arXiv preprint arXiv:NUMBER (2022).

[21] Daniel Kurt. 2021. The Side Effects of Bad Credit. https://www.investopedia.com/the-side-effects-of-bad-credit-4769783.

[22] Paul Marks. 2022. Algorithmic hiring needs a human face. Commun. ACM 65, 3 (2022), 17–19.

[23] Victoria McGeer. 2002. Developing trust. Philosophical Explorations 5, 1 (2002), 21–38.

[24] Victoria McGeer. 2008. Trust, hope and empowerment. Australasian Journal of Philosophy 86, 2 (2008), 237–254.

[25] Teodor Mihai Moldovan and Pieter Abbeel. 2012. Safe Exploration in Markov Decision Processes. In Proceedings of the International Conference on Machine Learning. 1451–1458.

[26] Onora O'Neill. 2002. Reith lectures 2002: a question of trust. Lecture 2: Trust and Terror. BBC Reith Lect (2002).

[27] Philip Pettit. 1995. The cunning of trust. Philosophy & Public Affairs 24, 3 (1995), 202–225.

[28] Devin G. Pope and Justin R. Sydnor. 2011. Implementing Anti-Discrimination Policies in Statistical Profiling Models. American Economic Journal: Economic Policy 3, 3 (2011), 206–231.

[29] Michael L Slepian and Daniel R Ames. 2016. 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, 2 (2016), 282–288.

[30] Skyler Speakman, Srihari Sridharan, and Isaac Markus. 2018. Three Population Covariate Shift for Mobile Phone-based Credit Scoring. In Proceedings of the ACM Conference on Computing and Sustainable Societies. 20.

[31] Lav R. Varshney and Kush R. Varshney. 2016. Decision making with quantized priors leads to discrimination. Proc. IEEE 105, 2 (2016), 241–255.

[32] Dennis Wei. 2021. Decision-Making Under Selective Labels: Optimal Finite-Domain Policies and Beyond. In Proceedings of the International Conference on Machine Learning. 11035–11046.

[33] Andrzej P. Wierzbicki. 1982. A Mathematical Basis for Satisficing Decision Making. Mathematical Modelling 3, 5 (1982), 391–405.

AUTHORS

Bran Knowles (b.h.knowles1@lancaster.ac.uk) is a Senior Lecturer in the Data Science Institute at Lancaster University, Lancaster, UK.
Jason D'Cruz (jdcruz@albany.edu) is an Associate Professor of Philosophy at University at Albany, SUNY, New York, USA.
John T. Richards (ajtr@us.ibm.com) is a Distinguished Research Scientist at IBM Research, T. J. Watson Research Center, New York, USA.
Kush R. Varshney (krvarshn@us.ibm.com) is a Distinguished Research Scientist at IBM Research, T. J. Watson Research Center, New York, USA.

SIDEBAR: STATISTICAL ILLUSTRATIONS OF HUMBLE AI

Skepticism

The absence of relevant features leads to uncertainty in the data and machine learning model. It also prevents valid causal modeling because the ignorability assumption of causal inference (all confounding variables available as features) is violated. In pursuing Humble AI, seeking out and including more informative features reduces uncertainty and better discriminates the untrustworthy from the trustworthy. As illustrated in Figure 2(a), this manifests as narrower likelihood functions that are better separated from each other and have smaller false negative and false positive rates (areas of the orange and blue regions) than Figure 1(a).

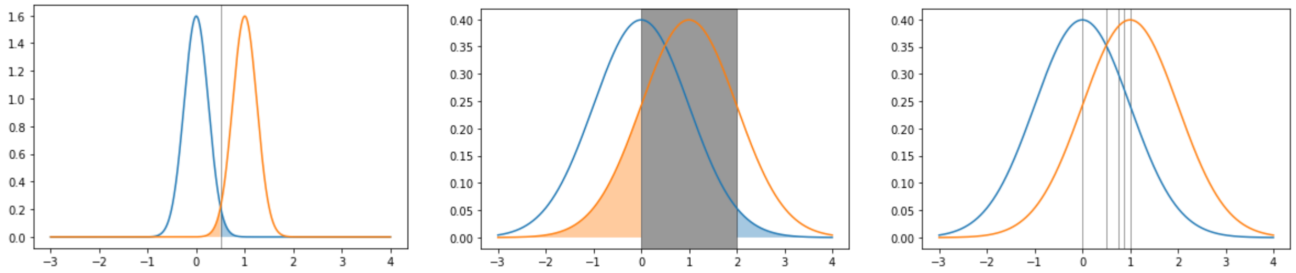
Curiosity

Curiosity about trust-responsiveness, e.g. in algorithmic hiring, might entail taking some portion of rejected candidates back into the pool and feeding data regarding their subsequent performance at interview (or if ultimately offered the job, their job performance) into the model to refine the AI's rejection criteria. The way to imagine such a process (illustrated in Figure 2(b)) is by expanding a decision threshold into a band—a sort of gray area—where the AI is most uncertain. Within this band, the trustworthy/untrustworthy determination is randomized. Notice that in Figure 2(b), the AI's false positive rate is the same as in Figure 1(b) (area of blue shading) with much smaller false negative rate (area of orange shading). Figure 2(c) illustrates how the curiosity-driven solution of [32] begins humbly and progressively moves the decision threshold from left to right.

Commitment

Related to the randomization method to achieving the curiosity stance of humble trust is making a commitment to "selective classification", also known as "classification with a reject option" [7]. In this paradigm, when the AI lacks confidence and is unsure whether a person is trustworthy or untrustworthy, it passes the decision on to a human decision maker (whose time and effort is costly). In practice, this amounts to the AI not classifying people who fall in a band around the decision threshold (this is the same kind of band used in randomizing the decision). Human decisions can also be fed back into model improvements, adjusting thresholds or providing hints of additional features of merit. Moreover, the existence of this human oversight can be made visible to those subject to AI decisions, in some cases involving a dialog between the human evaluator and the decision subject.

For paradigms such as selective classification to be tenable, it is critical that first, the AI system provides an indication of its confidence (this is known as uncertainty quantification [9]) and that second, the confidence is well-calibrated so that it is not over-confident or under-confident [10]. (Modern neural networks are notoriously over-confident [17].) Quantifying uncertainty is in itself a commitment to humility as is the provision of understandable explanations of a decision. Explanations of a negative decision, even explanations generated by the AI itself, can be cast as suggestive rather than definitive, and would ideally provide information about how the decision can be appealed or changed down the road through attention to one or more of the features that most contributed to the outcome. Finally, after deployment, ongoing monitoring can evaluate whether and why individuals are receiving negative decisions. This can expose areas of potential weakness in the model, supporting a continual process of improvement.



(a)

(b)

(c)

Figure 2: (a) Lowered uncertainty through more informative features. (b) The gray band around a decision threshold may be used for randomization or to revert to a human decision maker. (c) The decision threshold progressively moves from left to right, starting in a humble position.