Trustworthiness in the context of AI technologies

by Margaret Levi and the CASBS team

As the world changes, so does our understanding of what constitutes competence and accountability, two components of trustworthiness. For example, we expect airlines to provide us with a safe trip from point A to B. However, the original concern was that the plane not fall out of the air for mechanical reasons or because of exhausted or unskilled pilots, but now we also count on the airlines to protect us from terrorist actions, the hacking of their computers, and COVID. Sometimes, the institutions we create alter our expectations as we experience them. Government itself is a case in point; as government capacities and reach change, so do our demands on government and our fears of its power over us and others. Policing is an obvious example: As expectations of the police increase simultaneously with the police force becoming more militarized, both its tasks and its capacity to do both harm and good expand. The same could be said about government regulations in a whole variety of domains that affect our individual well-being, our national security, and the economic, social and environmental health of our societies.

AI faces a similar dynamic. As AI develops, it transforms our interactions with it and what we want of it and worry about. This means that the governance arrangements must keep pace with the combination of the transformations in the technology and the concerns of those affected by it. These technologies change the world as much as the transformation of government has. And, as with government, the forms of competence have to evolve as well.

In discussing whether we should “trust” AI, the same terms often carry distinctive meanings in different academic disciplines and for different industries. Although we are unlikely to achieve a common language, we can at least attempt to focus attention on the critical variables. Here our concern is not trust, that is when an individual perceives or judges another person or institution worthy of the risks involved with relying on them for a good or service.

What we really care about is *trustworthiness*, those qualities of organizations (e.g., governments, firms) meant to promote the public’s confidence in the intentions, competence, and accountability of the organization, the staff, their platforms, algorithms, and systems within specific domains. In other words, we are concerned less with trust as a subjective feeling that an individual might have about another person or institution; and more with trustworthiness as an objective property of the person or institution that might be the object of trust. *Intentions* refer to the extent that the relevant entity reveals an obligation to act with proper concern for those its actions and products will affect, be they investors, shareholders, consumers, or citizens. Competence refers to the capacities and ability of the entity to accomplish what it claims it can do. *Accountability* refers to the ways it takes responsibility for mistakes and unintended consequences of its actions and products.

Such qualities are generally created and sustained by mechanisms and institutions that humans design. Something or someone is objectively trustworthy because incentives are built into the governance arrangements in ways that ensure: 1) clarification of and delivery on obligations; 2) competent performance in terms of a particular end or task, and; 3) means for punishing offenders in the system and compensating stakeholders the system harms. More often than not, it is formal institutions such as laws, regulations, and management processes and rules that make credible the commitments to deliver on promises and competence. Reputational concerns, networks, and norms can often do some of the work. Whether external or internal, formal or informal, the incentives and sanctions help define the governance arrangements of the given organization. Whether or not individuals actually perceive these objective sources of trustworthiness or have confidence in them is a separate and important question but not the focus of this piece.

The many statements of AI ethics principles (Algorithm Watch has compiled 160) tend to define values, but very few speak about the mechanisms necessary to establish trustworthiness. Let's examine three areas of AI governance in the light of the above conception of trustworthiness. For each of these areas, we:

- Clarify the domain or area in which we expect trustworthiness and the obligations that entails (domain specificity - or trustworthiness to do what)
- Discuss means of providing confidence that the product can actually do what it is designed to do (competence)
- Elaborate some possible mechanisms for establishing credible commitments in relationship to obligations and promise-keeping (accountability)

1) Safety

Trustworthiness to do what: Achieve (or help achieve) the technology's stated goals without creating excessive safety risks for individuals or societies. Doing so can involve managing tradeoffs between beneficial and harmful outcomes. For example, semi-autonomous vehicles might prove capable of transporting people safely and efficiently in most scenarios, but they might also cause harm in unusual scenarios which most human drivers could navigate using common sense.

Competence: We know that unsuitable or inappropriate training data can create hazards. A medical diagnostic AI could make mis-diagnoses if trained on unsuitable data (see example from Michael Jordan here), or a fleet of autonomous vehicles could present excessive risks to passengers if there is a mismatch between the scenarios used to train the vehicles and the (potentially unfamiliar) scenarios in which the vehicles are deployed. But there are ways to improve competence. Two examples: First, the philosopher David Danks has proposed phased clinical trials for autonomous vehicles similar to the phased clinical trials overseen by the FDA. Second, there are precedents for government-approved safety organizations - such as Underwriters Labs - that provide scientifically rigorous inspection, auditing, and testing services to ensure product safety. It is conceivable that such arrangements could be extended to various AI technologies.

Accountability: Standards for accountability must be designed in. For example, if a pedestrian gets killed, when is the driver of the semi-autonomous vehicle responsible as opposed to the manufacturer? Presumably the possibility of class-action lawsuits, in addition to the possibility of long-term damage to brand reputation, are means of ensuring credible commitments.

2) Algorithmic fairness / lack of bias

Trustworthiness to do what: Achieve (or help achieve) the technology's stated goals without creating excessive procedural or distributive unfairness. For example, facial recognition or medical diagnostic technologies should presumably work comparably well across such dimensions as age, gender, ethnicity, skin tone, etc.

Competence: Some have advocated systems of algorithm auditing to ensure that the creators of algorithms are competent to avoid biases or appropriately manage and communicate tradeoffs between different concepts of algorithmic bias. An historical analog might be the state-level (in the US) regulation of personal insurance rating algorithms to ensure that insurance rates are fair and not excessive.

Accountability: Arrangements by which organizations can be fined, sued, or otherwise penalized for material violations of fairness. Damage to brand reputation.

3) Appropriate use of data

Trustworthiness to do what: Use the "digital exhaust" that people generate only in ways that users (a) consciously endorse, (b) improve the service they are receiving, and (c) do not harm the user or the larger society. Note that (a) would be violated if a click to accept were treated as a conscious endorsement when it clearly is not (for example when the document being agreed to is excessively lengthy or full of dense legal jargon). For (b), examples involve using click behaviors to suggest books, films, or other products using collaborative filtering; improving search algorithms; improving spell-check algorithms, etc. Examples of harm (c) include selling users' internet search data to a health insurer for pricing or underwriting purposes; or to a rogue organization such as in the Cambridge Analytica scandal.

Competence: The competence of an organization to prevent cyber-breaches or rogue elements either within or outside the organization to improperly gain access to sensitive data is a moving target, subject to technological and legal innovations. Perhaps regulators can appoint cyber-security inspection organizations (analogous to the UL arrangement mentioned under "Safety") to periodically assure the organization's competence along this dimension.

Accountability: Threats of lawsuits and damage to brand equity leading users to switch platforms are among the means for ensuring credible commitments. More important perhaps are government fines and the threat of being broken up or being regulated more strictly. And then there are a number of strategies under discussion addressing who owns the data.

These examples highlight the need for both technical and institutional mechanisms. The absence of such mechanisms has adversely affected the public on numerous occasions. For example, the recent algorithmically determined A-level scores controversy caused by the Covid-19 cancellation of the exams in the UK presents multiple areas where trustworthiness needs to be implemented. The students, their families, and school systems expect the government to provide a fair and trustworthy system and have the qualifications to implement scoring for university placement. Although delaying or holding socially distanced in-person exams were recommended, once the exams were canceled by the education minister, the Office of Qualifications and Examinations Regulation (Ofqual) had teachers estimate potential scores and algorithmically adjusted them according to the historic performance of secondary schools. Ofqual, therefore, had to count on teachers to provide a trustworthy assessment (for which they are well qualified to do so) and more tenuously, had to count on an algorithmic system to be trustworthy as well. But the system was not trustworthy. It downgraded nearly 40% of all exams. In response to mass protests by students and families whose lowered scores now reflected the historical class biases of the British secondary school system, the government apologized, the head of Ofqual resigned, and the scores were reverted to the original human recommended grades.

Dialogue on institutional and technical mechanisms has begun (Towards Trustworthy AI Development) and can help governments verify that these tools will do what they say they will do via third party auditing and model interpretability. These mechanisms and systems of contestability are especially important in sectors where the impacts of AI fall on more vulnerable populations such as AI used in surveillance or the criminal justice system.

Because the companies and platforms developing AI products cross so many borders and boundaries, international and multi-stakeholder collaboration on the governance of AI is urgently needed, if difficult. There are many multilateral efforts, e.g., the EU, Partnership on AI, OECD, and WEF. They all have and continue to develop institutional mechanisms to ensure their trustworthiness. These groups face many challenges such as holding their members accountable to their stated values without losing members. Coordinating groups also face power asymmetries among diverse parties such as human rights non-profits, big tech, and nations. These coordinating groups must also be trustworthy. Cooperation at higher levels of governance through treaties, trade agreements, and agreed-upon norms is especially important when it comes to AI uses such as lethal autonomous weapons. Verification of actions and some form of punishment for compliance failure is essential in these cases even when direct enforcement is not possible.

In some instances, inclusion in the design and implementation processes are conditions of trustworthiness in order that there is strong reason to believe that the voices of those affected are being taken into consideration from the start and throughout. "Nothing about us without us," a slogan of the disability movement, may be a good principle for AI as well.

Creating trustworthy AI also implies inclusion of the diverse concerns of the publics affected as well as educating the public about the risks of AI. This means providing means for consumers and publics to effectively assess the extent to which governments, companies, scientists, and relevant others are delivering on their promises regarding the benefits of AI, its quality, and its distribution.

We offer these illustrative questions to prime discussion.

1. What are the inducements to firms to enhance and publicize trustworthiness on issues and features where there is a conflict with profitability? Is brand reputation enough? Under what conditions are external enforcement mechanisms, including government regulations, necessary?
2. Might there be innovative business models in which the trustworthiness of a firm is considered integral to its success (e.g. the B Corporation model)?
3. If trustworthiness refers to objective characteristics and qualities of the person or organization one is then expected to trust, how do we ensure that the truster actually perceives those characteristics and understands them as a basis for confidence?
4. If the potential truster is skeptical of the good intentions of the organization, does that doom all efforts at establishing objective trustworthiness?

CENTER FOR ADVANCED STUDY IN THE BEHAVIORAL SCIENCES

The Center for Advanced Study in the Behavioral Sciences is a place where great minds confront the critical issues of our time, where boundaries and assumptions are challenged, where original interdisciplinary thinking is the norm, where extraordinary collaborations become possible, and where innovative ideas are in pursuit of intellectual breakthroughs that can shape our world. CASBS @ Stanford brings together deep thinkers from diverse disciplines and communities to advance understanding of the full range of human beliefs, behaviors, interactions, and institutions. A leading incubator of human-centered knowledge, CASBS facilitates collaborations across academia, policy, industry, civil society, and government to collectively design a better future.

Selected References

Cook, K., R. Hardin, et al. 2005. *Cooperation Without Trust?* New York, Russell Sage Foundation.

Hardin, R. 2002. *Trust and Trustworthiness*. New York, Russell Sage Foundation.

Levi, M. 2019. "Trustworthy Government, Legitimating Beliefs." In *NOMOS: LXI*, edited by J. Knight and M. Schwartzberg. New York, New York University Press.

O'Neill, O. 2002. *A Question of Trust*. New York, Cambridge University Press.