Toward a Scorecard (& Roadmap) for Trustworthy AI Implementations in Organizations

Munindar P. Singh & Roger C. Mayer
North Carolina State University

Abstract

The rapid expansion of Artificial Intelligence (AI) in organizations raises new concerns about how we might ensure that AI will work in the societal interest. Not only is AI sophisticated beyond previous technologies, it is also often an integral part of core business processes. An AI implementation cannot be properly understood nor its trustworthiness judged in isolation from the organizations and ecosystems in which it is applied. Accordingly, we propose a sociotechnical framework based on the well-known ability, benevolence, and integrity model of trust. We adopt four main concepts relating to what might be termed good behavior as recently identified by the US National Science Foundation as criteria to judge the trustworthiness of AI. These concepts are *Fairness*, *Ethics*, *Accountability*, *Transparency*. We provide working definitions of these concepts and develop a scorecard that stakeholders may use to assess the trustworthiness of an AI implementation.

Keywords

Assessing trustworthiness in AI; Critical questions for AI trust; Fairness; Ethics; Accountability; Transparency

1. Introduction

Artificial Intelligence (AI) is revolutionizing how business and other organizations carry out their internal processes and engage with customers, regulators, and other external stakeholders. Increasingly, AI is contributing to or even taking over major business functions. Manyika et al. (2019) review some opportunities and risks associated with AI. How can stakeholders make sure that AI will benefit society, helping both ordinary people and businesses in achieving their objectives?

Let us first lay out some important terminology. *AI* refers to any of a variety of technologies that demonstrate some form of automation of tasks that previously relied on human intelligence. An *AI agent* is a computational artifact that applies AI and is able to interact with humans, other agents, and information sources such as sensors. To relate agents to algorithms, we may think of an agent as an algorithm connected to sensors, effectors, and communication channels. A *sociotechnical system (STS)* refers to a conglomerate of social entities (its stakeholders: humans and organizations) and technical entities (AI agents and databases) working to achieve selected stakeholder objectives. An *AI implementation* (in an organization) is when AI technologies are deployed in the organization. For example, classification is an AI technology that can be implemented in various organizations: in a bank to classify loan applicants as credit-worthy and in a judicial court to classify a defendant as being at risk for recidivism.

The trustworthiness of AI and AI safety and ethics are related yet extant efforts have not explicitly captured their relationship. This chapter addresses how to do so by relating independently proposed elements of trust with elements of AI safety and ethics. It offers a scorecard based on critical questions to characterize the trustworthiness of AI implementations. This scorecard would help stakeholders, within or without in an organization, to critically evaluate AI implementations and guide their creation and continual improvement.

Specifically, we apply Mayer, Davis, & Schoorman's (1995) framework for trust (see Mayer & Mayer, this volume, for an overview). In brief, trust is the willingness of a trustor (trusting party) to make itself vulnerable to a trustee (party to be trusted) for a particular task or purpose when the trustee cannot be monitored or controlled. Whereas trust is an intention (i.e., a willingness) held by the trustor, trustworthiness is the trustor's evaluation of the trustee. The important outcome of trust is the trustor actually engaging in behavior that puts the trustor at risk. Trustworthiness consists of three dimensions: the trustor's perceptions of the trustee's ability, benevolence, and integrity (ABI). *Ability* concerns how proficient the trustor deems the trustee to be at the task at hand. *Benevolence* concerns how strongly the trustor expects the trustee to seek the benefit of the trustor, i.e., to support the trustor's interests. *Integrity* concerns how consistently the trustor believes the trustee will follow values acceptable to the trustor. Thus, this approach separates perceived characteristics of the trustee (trustworthiness) from trust (a behavioral intention) and risk-taking behavior that makes the trustor vulnerable to the trustee.

Kaplan, Kessler, Brill, & Hancock's (2020) meta-analysis including 65 articles on trust in AI defines trust as "the *reliance* by an agent that actions *prejudicial* to their well-being will not be undertaken by influential others" (emphasis added). This definition limits consideration of trust in two ways. Firstly, reliance is influenced by factors other than trust, such as a lack of viable alternatives. Moreover, reliance does not provide any diagnostic information about why the trustor takes the risk of relying on the technology. Consideration of trust as a willingness to be vulnerable and trustworthiness as comprising three perceptions of ability, benevolence, and integrity provides more insight into the reasons the trustor made themself vulnerable to the trustee. Secondly, in terms of the ABI model, if a trustee has a prejudice toward the trustor, it would be reflected in benevolence. A lack of prejudice is a neutral orientation, which falls short of a perceived desire to do things to benefit the trustor, captured by the higher end of the benevolence scale. In addition to the trustee's orientation toward the trustor, ability and integrity as described above are focused on distinct issues. Thus, use of the three dimensions of trustworthiness enables us to consider a broader spectrum of reasons that might influence a trustor to either be willing or to avoid being willing to be vulnerable to the AI. We therefore adopt the ABI model for its diagnostic utility to delineate why a trustor might trust an AI implementation to a greater or lesser extent.

Contributions in a Nutshell

The US National Science Foundation (NSF) is interested in research considering the impact of AI on people. For this purpose, the NSF recently introduced four dimensions of AI to be considered: fairness, ethics, accountability, and transparency (FEAT). [https://www.nsf.gov/pubs/2019/nsf19016/nsf19016.jsp]. They intentionally did not define these terms, encouraging instead that researchers define them as they deem warranted. We offer

working definitions for these terms that help us highlight our main contributions, namely, a scorecard based on answers to critical questions.

We use both ABI and FEAT as bases to evaluate AI implementations in organizations. Doing so leads us to identify questions in the resulting 3×4 grid. We posit that answering these questions will provide a fruitful basis for evaluating AI implementations in terms of an AI system's trustworthiness in the eyes of those affected by it. We close with a discussion of some promising directions for investigation.

2. A Sociotechnical Stance on AI

> Any sufficiently advanced technology is indistinguishable from magic.
> —Arthur C. Clarke (1968)

In common usage, AI is framed as a set of mysterious artifacts that can magically solve problems. To end users, the often uncanny ability of AI to identify patterns and make predictions is nothing short of magical, echoing Arthur C. Clarke's famous dictum. Therefore, it is perhaps not surprising that they focus on the technical aspects of AI. Indeed, from the perspective of trust, that view is not entirely without merit because the ability associated with or ascribed to AI relies on its construction and function as a technical artifact. Viewing AI as a purely technical artifact makes it more difficult to ascribe integrity and particularly benevolence to it.

A second view of AI is through the so-called *intentional stance* (Dennett 1987; McCarthy 1979). The idea of the intentional stance is that we (as humans) may ascribe a mind to any technical artifact (e.g., see Wingert & Mayer, this volume). Instead of seeking to understand the artifact in terms of its design or function, we would ascribe a mental state to it, usually in terms of the so-called folk psychological concepts such as beliefs, knowledge, goals, and intentions to describe, understand, and explain its behaviors. In a famous example, we might view an old-fashioned (that is, not infused with AI in the modern sense) thermostat in intentional terms. We might state that the thermostat intends to raise the temperature to at least the set point. When it believes the temperature is below its set point, in accordance with its intention, it would turn the central furnace on to bring the temperature up. We might be able to explain when the thermostat malfunctions, for example, by determining that either its belief is wrong (the temperature is high enough so its sensor may be broken) or that its intention (reflecting raising the temperature) may be out of sync with the user's. While the intentional stance remains the dominant view of AI within computer science, another influential formulation of it is as the *knowledge level* (Newell 1982, 1993).

This stance is seen in recent work on the ethics of AI, for example, in building AI agents that demonstrate ethical reasoning (e.g., Bremner et al. 2000). An important benefit of this stance is that it can help combat complexity. Instead of having to contend with the incredible complexity of the construction of today's computing artifacts, we can form a rough-and-ready model of them in folk psychological terms—terms that are familiar to us as humans and help to mediate our normal interactions with other humans.

Because of its ascription of mental states to artifacts, the intentional stance makes it quite natural to talk of the benevolence and integrity of the artifacts. However, for our present purposes, such ascriptions can be misleading in that they hide the contributions of social actors, such as humans and organizations, in how AI agents interact with people. Specifically, although the intentional stance is couched in psychological language, it is very much a view of AI as a technical artifact—that is, divorced from its societal or organizational contexts. Moreover, the mysterious nature of AI as perceived by lay people is arguably made more prominent through the use of such psychological language.

In light of the foregoing, in this chapter we consider AI implementations as part of a sociotechnical framework. The idea of a sociotechnical system was developed in Trist & Bamforth's (1951) famous studies of coal miners, referring to the combined relationship of human and technical aspects of a workplace. For our purposes, the AI is not a standalone artifact, whether or not we ascribe a mental state to it. This view is developed further in Singh (2013, 2022). AI as implemented in an organization therefore reflects, or ought to reflect, the purpose and goals of the organization. AI is often packaged with Big Data (technologies for acquiring, storing, querying, and manipulating vast amounts of data). In such cases, the sociotechnical system must include the contexts in which the data are obtained and curated.

An AI implementation's ability depends upon how effectively it meets its organizational purpose, for example, to serve the organization's stakeholders and solve their problems within a given domain. Its benevolence arises (or not) from not just the AI technology, but how well the technology in combination with the organizational backdrop takes its stakeholders' interests into account. Likewise, the sociotechnical system's integrity arises (or not) from how the technology and organization together respect societal and legal norms.

3. Ability, Benevolence, and Integrity

We employ the ABI framework of trust from Mayer, Davis, & Schoorman (1995), introduced above, focusing here on how it applies to evaluation of an AI implementation.

Ceteris paribus, if the AI implementation has stronger ability for the task at hand, then the chances of a good outcome are better; if the AI implementation lacks ability (e.g., poor sensors, poor data, or poor algorithms) to do what is needed, then it is likely unwise to make oneself vulnerable to that trustee.

Benevolence addresses the question of to what extent the AI implementation will seek to protect the interests of the trustor. In many organizational settings where AI is implemented, a trustor may have no personal relationship with the organization and definitely not with its backroom AI implementation. For example, a loan applicant would often have little more than a transactional relationship with their bank. Here, we adopt Hamm, Smidt, & Mayer's (2020) idea from their study of trust in the federal government, to understand benevolence as the perception of the extent to which the trustee cares about oneself *and others similar to oneself*. Hamm et al. provide empirical evidence in support of this understanding, including an individual's willingness to be vulnerable to the trustee (the government in their study) and the conception providing higher correlation with behavior than the traditional American National Election Studies (ANES)

measure that has been used in political science for decades. For this reason, we formulate the benevolence of an AI implementation from a perspective of its holding the interests of *the trustor and others like the trustor*.

An assessment of integrity necessitates two underlying assumptions. The first is that the trustor holds an acceptable set of values. If, for example, an AI implementation were structured around the goal of short-term gain for the organization, to the extent that this was inconsistent with the trustor's expectations of what the organization ought to do, this would diminish the AI implementation's integrity. In addition, if it did not reliably and consistently follow the set of values it purports to adhere to, that would also reduce its perceived integrity.

It is important at this point to make two notes about this use of the ABI model. Firstly, the theory was originally described using the language of "parties" to denote individuals. Soon afterwards, its authors explained that the model was intentionally designed to be isomorphic (Schoorman, Mayer, & Davis, 1996), meaning that the definition and conceptualization is the same across different levels of analysis (e.g., Rousseau, 1985). The constructs of trust and the three trustworthiness factors are applicable not only to interpersonal trust, but also to intergroup and interorganizational trust, and trust between these levels of analysis. We therefore posit that it is appropriate to apply the model to a sociotechnical system, as it is simply a technologically enhanced system of people.

Secondly, we note here a distinction between integrity and benevolence. While integrity is the perception that the trustee adheres to an acceptable set of values, benevolence reflects a perception of the relationship between the trustee and the trustor. In some cases, benevolence and integrity are aligned, possibly to the point of being indistinguishable from one another. However, in many situations, judgments of benevolence and integrity may diverge so it is fruitful to consider these as separate factors. Consider, for example, that a bank has implemented AI for mortgage loan decisions that favors married over single applicants. A married couple who thought single people ought not to be discriminated against may recognize the AI implementation as having low integrity but high benevolence toward them. Conversely, an AI implementation that promoted diversity by biasing its positive mortgage decisions toward underrepresented groups may be seen by someone of a non-underrepresented group as having high integrity but low benevolence toward them.

4. Understanding Fairness, Ethics, Accountability, and Transparency

The NSF did not define fairness, ethics, accountability, and transparency. These terms have varying definitions in the literature and, indeed, the NSF's selection of terms is arguably idiosyncratic. However, they do address important intuitions about ensuring AI serves stakeholder needs and help us make our contribution. We circumvent endless debate on these terms by providing working definitions that capture key intuitions. We interpret Fairness in terms of how distributions of desirable outcomes are made, such as the parity in decision making across demographic groups. Our use of Ethics is in terms of deserts, or the extent to which the AI makes decisions that favor those who deserve the favor. We understand Accountability in terms of how account-giving is provided along with course correction of decision making. Finally, we

consider Transparency as being about the clarity of (including explanations provided about) the decisions made by the AI.

5. ABI on FEAT: Toward a Scorecard

The combinations of each of the four FEAT factors described above with each of the three factors of trustworthiness are described in Table 1—whose rows are the ABI factors in trust and whose columns are the FEAT factors. As a running example, consider an AI implementation that helps a bank decide mortgage loans.

To motivate our scorecard, we turn to the notion of a *critical question*, as proposed by Walton, Reed, & Macagno (2008). A critical question raises a concern at the heart of the robustness of an argument—that is, answering a critical question helps complete an argument. Critical questions, in essence, reflect knowledge of important concerns that an expert can raise in critically evaluating a claim made by another person or in formulating a defensible claim of their own. For example, for practical reasoning (as to select an action), the relevant critical questions would concern the feasibility of the action or the relationship between that action and one's goals.

Accordingly, in our present setting, we identify critical questions pertaining to the trustworthiness of AI as deployed in an organization. We place these questions in the respective cells of our table, focusing on the cell's particular line and column. We further use these critical questions as a basis for a reasoned scorecard, wherein actions for the critical questions can be summed across rows or columns to arrive at a measure of trustworthiness of an AI deployment in an organization.

Table 1
Critical Questions for Understanding the Trustworthiness of AI

	Fairness	Ethics	Accountability	Transparency
Working definition in brief	Distributions of outcomes, e.g., parity in decision making across groups	Just deserts, or the extent to which the AI makes decisions that favor deserving candidates	Ability to demand and provide justifications and make course corrections	Clarity and intelligibility of decisions
Ability	Does the AI possess or can it access the data and hardware (e.g., sensors) to make equitable decisions? Is its programming sophisticated enough to enable it to do so?	Can the AI incorporate data in its decision making that enables it to make just decisions; e.g., considering the personal situation of an application (children, credit history)? Does the AI have the capacity to incorporate feedback from external sources to improve the fairness of its decisions?	Does the AI assemble the data to justify its decisions to regulatory bodies and to prospectively take regulatory guidance about its decisions (e.g., loan approvals)?	Does the AI reveal its decision making, e.g., through explicit criteria under which a decision (e.g., loan approval or denial) is made? Can it do it in a way that is understandable to stakeholders?
Benevolence	Is the AI programmed to learn to better help me or people like me?	Is the AI structured to help individual applicants, e.g., by finding mitigating circumstances for their credit lapses or acknowledging that as minority or immigrant applicants they may not have the family backing to co-sign a loan?	Does the AI seek to make any necessary corrections in light of any problems discovered?	Does the AI reveal elements of its decision making and data to help the loan applicant and not, e.g., to mislead them? Does it help them be more successful in the future?
Integrity	Does the AI compute the fairness criteria honestly, e.g., not altering the	Does the AI avoid misusing information obtained from applicants to hurt their prospects, especially optional information?	Does the AI provide truthful justifications for its decisions? Does it make a concerted effort	Does the AI reveal elements of its decision making and data clearly and completely?

	criteria or ranges over which the demographic distributions are computed?	Does it match its stakeholders' moral expectations? What moral compass should the AI be built around?	to align itself better with stakeholders' values?	

6. Embodied AI: Robots and Vehicles

By embodied AI, we refer to AI realized in physical artifacts such as autonomous robots and vehicles. Glickson & Wooley (2020) refer to this as tangibility. More generally, we can think of AI realized in cyber-physical systems, which would include chemical refineries and health care devices such as insulin pumps and cardiac pacemakers. Glickson & Woolley observe that tangibility increases trust in AI. It is not clear how this effect varies with tangible artifacts that may not be fully apparent in a user's consciousness. However, we expect that questions of the sort we developed above would be definable for these settings as well.

When dealing with a physical artifact of any scale that acts intelligently, it is difficult not to conceive of the AI in it as "the ghost in the machine" in Gilbert Ryle's memorable phrase (Ryle, 1949). However, a small amount of reflection brings forth the sociotechnical nature of even such seemingly standalone artifacts.

For embodied AI, we can identify important organizational roles such as builder (e.g., see Wingert & Mayer, this volume), owner, operator, and maintainer. In general, these roles may be filled by distinct organizations. Let's consider the case of present-day vehicles, that is, those not infused with AI technology. The builder of a vehicle would be an automobile manufacturer such as Chrysler. Its owner may be a leasing company such as Hertz, its operator may be the person who rents the vehicle, and its maintainer may be a franchisee of or contractor for Hertz. The same roles would apply to autonomous vehicles. Therefore, when we apply our framework to embodied AI, we would likewise need to consider the FEAT criteria with respect to these organizational roles (van der Werff, Blomqvist, & Koskinen, 2021). For example, we might refine the scorecard by addressing the needs of both *internal* (those who constitute the organization and are primarily responsible for carrying out its processes) and *external* (those whom the organization serves or otherwise affects) stakeholders.

7. Conclusions and Future Directions

We can view Table 1 as an outline for a research roadmap to develop AI tools and implement them in organizations in a manner that would support positive answers to the questions raised above. Alternatively, we can map each question from "does" to "how does" to produce challenges for focusing AI technology development on improving trustworthiness. Both of the above directions would require research in understanding what stakeholders need and value. In general, stakeholders cannot express and may not know what these are and an iterative approach is needed to continually refine an AI implementation to keep it aligned with their needs and

values. Such research would need to be interdisciplinary, going beyond (1) current computing practice to produce AI implementations that consider the societal context in interpreting stakeholder needs and values and making major course corrections (not merely minor personalization) and (2) management science to produce methods that keep up rapid behavior shifts of the underlying technology.

One inherent complexity that deserves consideration is that the three dimensions of trustworthiness can interact. We described them in this paper ceteris paribus, or in their simple form and in isolation. For example, considered alone, greater ability should serve to make AI more trustworthy. But what if the AI is structured not to attempt to be benevolent towards a stakeholder, but is instead seen by the trustor as malevolent? If the AI is set up to catch violations of some rule or law, then for those stakeholders who deem that they may be damaged by the AI's actions, the higher the AI's ability the *less* they would judge the AI to be trustworthy. In fact, as AI is developed with increasing machine learning capabilities, those stakeholders would not only tend to avoid interaction with the AI, but would increasingly expend resources to nullify the AI's capability to execute its mission.

Recall the science fiction thriller movie "2001 A Space Odyssey" (with HAL, a computer gone mad) and the "Terminator" movie series (with Skynet, a self-aware but malevolent AI). These fictional AI agents illustrate the condition wherein greater AI ability may well lead to decreased trust. Future research should seek a deeper understanding of judgments of AI's trustworthiness by considering the interactions among the AI's ability, benevolence, and integrity.

We focused here on the stakeholders being society, or end users of the AI sociotechnical systems. Importantly, Lockey & Gillespie (this volume) clarify the importance of considering multiple stakeholders as trustors. Situations of particular subtlety are those where being trustworthy to one stakeholder conflicts with being trustworthy to another; Singh & Singh (2022) review legal precedent with potential applicability to trustworthy AI. We encourage further consideration of other stakeholders, which was beyond the scope of the present chapter.

Using the lens of interactions among the trustworthiness dimensions for trust in AI, consider the use of robots to lead people to safety as described by Allen Wagner (this volume). How would the ability of a robot designed to lead people to safety affect trust in the robot if the to-be-saved stakeholders were either library patrons, or maximum security prison inmates? Would the inmates consider the AI's benevolence to be as high as in the case of the library patrons? Would a presumed lower level of perceived AI benevolence make a more sophisticated robot less trustworthy to prisoners than would be a less sophisticated robot?

Per our opening comments, it can be expected that AI will play an increasingly important role in society. We hope this chapter leads researchers to formulate better questions concerning the trustworthiness of this technology.

Acknowledgments

We thank Nicole Gillespie and Oliver Schilke for helpful comments on an earlier version. MPS thanks the US National Science Foundation for support under grant IIS-2116751.

Biosketch

Munindar P. Singh is an Alumni Distinguished Graduate Professor in Computer Science at NC State University. His interests include the engineering and governance of sociotechnical systems and the ethics and safety of AI. Singh is a Fellow of AAAI, AAAS, ACM, and IEEE and a former Editor-in-Chief of *IEEE Internet Computing* and *ACM Transactions on Internet Technology*. Contact him at singh@ncsu.edu.

References

Bremner, P., Dennis, L. A., Fisher, M., and Winfield, A. F. T. (2019). On proactive, transparent, and verifiable ethical reasoning for robots. Proceedings of the IEEE, 107(3):541–561. DOI: 10.1109/JPROC.2019.2898267

Clarke, A. C. (1968). Clarke's Third Law on UFOs. Science., volume 159, number 3812, p. 255. DOI: 10.1126/science.159.3812.255-b

Dennett, D. C. (1987). The Intentional Stance. Cambridge, Massachusetts: MIT Press.

Glickson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Annals of the Academy of Management, 14:2, 627-660. DOI: 10.5465/annals.2018.0057

Hamm, J., Smidt, C., & Mayer, R. C. (2019). Understanding the Psychological Nature and Mechanisms of Political Trust. PLOS ONE, first published online May 15, 2019. DOI: 10.1371/journal.pone.0215835

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. Human Factors, DOI: 10.1177/00187208211013988

Manyika, J., Silberg J. & Presten B. (2019). What Do We Do About the Biases in AI? Oct 25, 2019. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. The Academy of Management Review, 20(3):709–734. DOI: 10.5465/amr.1995.9508080335

McCarthy, J. (1979). Ascribing mental qualities to machines. In Ringle, M., editor, *Philosophical Perspectives in Artificial Intelligence*, pages 161–195. Humanities Press, Brighton, UK.

Newell, A. (1982). The knowledge level. Artificial Intelligence, 18(1):87–127. DOI: 10.1016/0004-3702(82)90012-1

Newell, A. (1993). Reflections on the knowledge level. Artificial Intelligence, 59(1):31–38. DOI: 10.1016/0004-3702(93)90166-9

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. Research in Organizational Behavior, 7, 1-37.

Ryle, G. (1949). The Concept of Mind. Oxford, UK: Oxford University Press.

Singh, M. P. (2013). Norms as a basis for governing sociotechnical systems. ACM Transactions on Intelligent Systems and Technology (TIST), 5(1):21:1–21:23. DOI: 10.1145/2542182.2542203

Singh, M. P. (2022). Consent as a foundation for responsible autonomy. Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI), 36(11):12301–12306. Blue Sky Track. DOI: 10.1609/aaai.v36i11.21494

Singh, A. M. and Singh, M. P. (2023). Wasabi: A conceptual model for trustworthy AI. IEEE Computer, 56(2). In press. DOI: 10.1109/MC.2022.3212022

Van der Werff, L., Blomqvist, K., & Koskinen, S. (2021). Trust Cues in Artificial Intelligence: A Multilevel Case Study in a Service Organization. In Gillespie, N., Fulmer, C. A. R., & Lewicki, J. editors, Understanding Trust in Organizations: A Multilevel Perspective. Routledge.

Walton, D. N., Reed, C., and Macagno, F. (2008). Argumentation Schemes. Cambridge University Press, Cambridge, UK.