# Requirements for Trustworthy AI

Subjects: Computer Science, Artificial Intelligence

Contributor: Michael Mayowa Farayola , Irina Tal , Regina Connolly , Takfarinas Saber , Malika Bendechache

Artificial Intelligence (AI) can be very beneficial in the criminal justice system for predicting the risk of recidivism. AI provides unrivalled high computing power, speed, and accuracy; all harnessed to strengthen the efficiency in predicting convicted individuals who may be on the verge of recommitting a crime. The application of AI models for predicting recidivism has brought positive effects by minimizing the possible re-occurrence of crime. However, the question remains of whether criminal justice system stakeholders can trust AI systems regarding fairness, transparency, privacy and data protection, consistency, societal well-being, and accountability when predicting convicted individuals' possible risk of recidivism. These are all requirements for a trustworthy AI.

trustworthy AI criminal justice system trust

## 1. Trust

Trust is a crucial component of success for risk assessment tools for recidivism to thrive [1][2]. Trust is a complex area of focus that has drawn the attention of many practitioners and scholars across different disciplinary fields [3]. This attention has focused on understanding the antecedents of trust, the conceptualization of trust, forms or types of trust, people involved in trust, and how trust impacts our ethics and the associations at different levels of life. Despite this, there is a surprising lack of consensus regarding how trust should be defined. Nevertheless, researchers provide some definitions and fundamental concepts of trust across several domains, as it will serve as a prerequisite to what trustworthy AI entails.

One overview of trust in commercial and personal transactions in the digital age [4] describes trust as an interpersonal phenomenon that facilitates human relationships by reducing uncertainty risks. Trust is the confidence level a trustor has in a trustee to do the right things. Trust can be attitudinal, predictable, and voluntarist. The attitudinal view of trust focuses on a trusting attitude due to personal beliefs. The predictability view of trust focuses on the notion of the positive expectation of a trustor, based on the trustee's behavior, that the trustee will act benevolently. Lastly, the voluntarist view of trust is the state of voluntary subjection of a trustor to the vulnerability of a trustee, that is, a position of risk. Therefore, one can affirmatively say that trusting involves risk. A study on public trust in local government, explaining the role of good governance practices [5], defines trust as a psychological state that constitutes a willingness to take risks based upon positive expectations of a trustee's behavior. Therefore, trust is a bridge between a trustor and a trustee and a lubricating factor for a consistent relationship. However, to strengthen trust, there is a need for transparency, accountability, and responsiveness on the part of both parties [6].

More on the definition of trust, a discussion on the significance of trust for organizational accountability [7] defined trust as holding a trustee accountable over time with the notion that they will exhibit integrity and honesty. Trust is conceptualized as mutual expectations and reciprocity between the trustor and trustee, strengthening social interaction. Trust also involves the reduction of complexity, ethical accountability, and responsibility. Furthermore, trust between different parties is instituted based on standard norms. In essence, trust can act as a sense of accountability on the part of a trustee and the state of a trustor's vulnerability. In conclusion, a review on trustworthy AI [8] defined trust as the willingness of a trustor to depend on a trustee due to a lack of control over the trustee, thereby making available the opportunistic behavior of the trustee.

It is worth noting that trust develops over time based on the trustee's behavior and conformation to a trustor's beliefs. Therefore, it is essential to understand what precedes the establishment of trust, referred to as the antecedent of trust [4][5][9][10]. The antecedents of trust are ability, benevolence, integrity, and predictability. Ability refers to a trustee's skills, characteristics, and level of competence in a specific domain. Benevolence is a trustee's impulse willingness to do good to a trustor, putting aside self-gain profit. Integrity is the perception that the trustee will always act with consistent and positive values. Predictability assures trust will be sustained throughout the relationship between parties. In all, ability, benevolence, integrity, and predictability can be bracketed as attributes used in judging the trustworthiness of a trustee by a trustor.

Now that we have established the basic concept of trust, the question is, what is the relation between people, trust, and technology, such as AI systems? The relationship is such that people and societies are the trustors, the AI system is the trustee, and the connecting bridge is the trust. For people and societies to trust AI systems and subject themselves to a position of risk and vulnerability, the system needs to be trustworthy. Another question now is what trustworthiness is. Trustworthiness is being competent and committed to doing and achieving the expectation of a trustor, and trustworthiness is regarded as a virtue possessed by a trusted party [11]. The guidelines set up by the European Commission (EC) [12] defined trustworthiness as a prerequisite need for stakeholders(i.e., people affected by AI systems) to develop, deploy and use AI systems. This scenario led to what is now commonly called Trustworthy AI.

Before diving into what trustworthy AI entails, there is a need to understand other qualities that influence stakeholders' trust in AI systems. There are three qualities: human, environmental, and technological qualities [9]. Human qualities or attributes are associated with unique cultural backgrounds, past experiences, and ideologies. These qualities determine the extent to which an individual will voluntarily be subject to a state of risk at the expense of the trustee's freedom. Environmental qualities entail elements that propagate the level of trust in the deployed environment of AI technologies. These elements include the environment's cultural background, educational system, environmental awareness, technological advancement level, and technology tasks. Technical qualities focus on efficiency in yielding results, conformation to a level of expected performance, and processes in achieving outcomes. In [8], trust in technology is further classified based on the technology functionality, helpfulness, reliability, predictability, performance, purpose, and process.

## 2. Trustworthy AI

The benefits of AI in different spheres of life cannot be overemphasized. However, different conditions necessitate AI systems to be considered trustworthy. Several issues are related to developing and deploying AI models, such as violating individual privacy, racial bias, misunderstanding of its processes, and decision-making.

Trustworthy AI encapsulates the must-have qualities of the AI that warrant ethical approaches [9]. A review of trustworthy AI states that incorporating trust in AI's development and design will enable stakeholders to fully realize its potential [8][12]. A study [13] defined trustworthy AI as fair, secure, robust, transparent, safe, and explainable systems regarding human privacy and fundamental rights, and stakeholders involved in its development, deployment, and use are accountable.

In 2019, the European Commission (EC) developed "The Ethics Guideline for Trustworthy AI [12][14]". According to EC guidelines, trustworthy AI should be ethical, lawful, and robust, creating a foundation for stakeholders to trust AI systems' development, deployment, and usage. The guidelines provided four Ethical principles (i.e., respect for human autonomy, prevention of harm, fairness, and explicability) and a list of seven requirements for trustworthy AI (i.e., human agency and oversight, technical robustness and safety, privacy and data governance, transparency, non-discrimination and fairness, societal and environmental wellbeing, and accountability). A point to note is that these seven requirements are non-exhaustive, meaning several other requirements still apply to different domains. Still, these seven can serve as the base for any public or private sector considering trustworthy AI in its activities [12].

## 3. An Overview of the Seven Requirements for Trustworthy AI Proposed by the European Commission

The following discussed requirements of trustworthy AI are the requirements proposed by the EC that can be applied and serve as the basis for any public or private domain. For every field, other requirements need to be considered and added to the proposed seven requirements by the European Commission to actualize trustworthy AI in such fields.

### 3.1. Human Agency and Oversight

Human agency and oversight revolve around fundamental human rights, human agencies, and the human administration of AI systems. AI systems should be built to support human autonomy in decision-making and not infringe on their fundamental rights. AI users should be at liberty to make decisions without AI systems making an adverse impact. This act will give AI users a sense of responsibility and freedom and enable trust in AI technology. AI systems should provide the required information to their users to better understand and interact with the system, allowing users to challenge the AI system's decisions when needs arise. Lastly, humans should engage in the decision process (human-in-the-loop), design cycle (human-on-the-loop), and the overall activities of the AI system (human in command).

### 3.2. Technical Robustness and Safety

AI systems are beneficial to the human race. However, if proper mechanisms are not in place, AI systems can cause harm. AI systems should bring safety to their users and prevent harm in every possible instance. AI systems contain data information, and an attack on the system can influence its outcomes leading to biases or harm to society. AI systems must be secure and built to withstand external attacks or threats. In adverse situations, AI systems should have a fallback plan to safeguard users and data information.

## 3.3. Diversity, Non-Discrimination, Fairness

Fairness is a requirement that has received the focus of many AI stakeholders since the advent of AI systems. Fairness must constantly be taken into account while developing AI systems. AI systems are vulnerable to prejudice if the AI development design lacks appropriate bias mitigation techniques. Hence, the developers should ensure the development of the AI system is void of discrimination and bias. In conclusion, AI systems should be inclusive and accessible to all social groups irrespective of their demographic information.

## 3.4. Accountability

Accountability is a requirement that enables the trust of AI stakeholders in AI systems when there is a level of responsibility and answerability. AI systems are inanimate tools, interacting with humans and influencing the decisions of their users directly or indirectly. Organizations should bear full responsibility in cases of negative impact caused by AI systems at different user instances. Users trust the system more when there is a level of answerability for its decision-making, especially with adverse effects.

## 3.5. Transparency

Transparency entails deliberate documentation, detailing, and understanding of AI systems, such as the data collection processes, design processes, and purpose of building such AI systems. When understanding the system’s underlying structure, transparency enables smooth auditing of AI systems. In essence, the procedures followed throughout the AI system design should be well-documented and answerable for issues related to the AI system. Transparency gives a head start on why AI systems behave in a particular manner and produce its outcome.

## 3.6. Privacy and Data Governance

Data are crucial in the development of AI systems. Apart from the models used, AI systems mainly function based on the consumption of large datasets used in designing the system. It is unarguably vital for developers to put in efforts toward the quality of data used for developing AI systems. Data are a significant source of biases in AI systems. Data tend to be biased without mitigating procedures to curtail bias. In addition, data privacy is paramount in the development of trustworthy AI. Access to the data should be restricted to authorized personnel only.

## 3.7. Societal and Environmental Wellbeing

AI systems have numerous benefits and have come to stay. Developers must build AI systems in such a way that they do not cause harm to humans. They should be designed to enhance humans' capabilities and not impose on their fundamental rights. AI systems should be eco-friendly, sustainable, and maintained. Lastly, developers and authorized stakeholders should oversee the AI system at all times to detect possible adverse effects it may cause to its users and the environment.

## References

1. O'Loughlin, T.; Bukowitz, R. A new approach toward social licensing of data analytics in the public sector. Aust. J. Soc. Issues 2021, 56, 198–212.
2. Alikhademi, K.; Drobina, E.; Prioleau, D.; Richardson, B.; Purves, D.; Gilbert, J.E. A review of predictive policing from the perspective of fairness. Artif. Intell. Law 2021, 7, 1–17.
3. Ryan, M. In AI we trust: Ethics, artificial intelligence, and reliability. Sci. Eng. Ethics 2020, 26, 2749–2767.
4. Connolly, R. Trust in commercial and personal transactions in the digital age. In The Oxford Handbook of Internet Studies; Oxford University Press: Oxford, UK, 2013; pp. 262–282.
5. Beshi, T.D.; Kaur, R. Public trust in local government: Explaining the role of good governance practices. Public Organ. Rev. 2020, 20, 337–350.
6. Smit, D.; Eybers, S.; Smith, J. A Data Analytics Organisation's Perspective on Trust and AI Adoption. In Proceedings of the Southern African Conference for Artificial Intelligence Research, Virtual, 6–10 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; Volume 1551, pp. 47–60.
7. Rendtorff, J.D. The significance of trust for organizational accountability: The legacy of Karl Polanyi. In Proceedings of the 3rd Emes-Polanyi Selected Conference Papers, Roskilde, Denmark, 16–17 April 2018; Roskilde University: Roskilde, Denmark, 2018.
8. Thiebes, S.; Lins, S.; Sunyaev, A. Trustworthy artificial intelligence. Electron. Mark. 2021, 31, 447–464.
9. Toreini, E.; Aitken, M.; Coopamootoo, K.; Elliott, K.; Zelaya, C.G.; Van Moorsel, A. The relationship between trust in AI and trustworthy machine learning technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 272–283.
10. Liu, K.; Tao, D. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. Comput. Hum. Behav. 2022, 127, 107026.

11. Sutrop, M. Should we trust artificial intelligence? Trames A J. Humanit. Soc. Sci. 2019, 23, 499–522.
12. High-Level Expert Group on Artificial Intelligence. In Ethics Guidelines for Trustworthy AI; European Commission: Brussels, Belgium, 2019. Available online: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (accessed on 3 July 2023).
13. OECD. Tools for Trustworthy AI: A Framework to Compare Implementation Tools for Trustworthy AI Systems; OECD Digital Economy Papers, No. 312; OECD Publishing: Paris, France, 2021.
14. Floridi, L. Establishing the rules for building trustworthy AI. Nat. Mach. Intell. 2019, 1, 261–262.

Retrieved from https://www.encyclopedia.pub/entry/history/show/107446