Towards Trustworthy AI: Blockchain-based Architecture

Vipul Popat   
*School of Computing*  
*Dublin City university*Galway, Republic of Ireland   
vipul.popat2@mail.dcu.ie, 0000-0002-0511-4563

Vishal Padwal  
*School of Computing*  
*Dublin City university*Dublin, Republic of Ireland   
Vishal.padwal2@mail.dcu.ie

Dr. Irina Tal  
*Faculty of Engineering and Computing*  
*Dublin City university*Dublin, Republic of Ireland   
irina.tal@mail.dcu.ie

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Introduction

Blockchain technology can play a role in enabling Trustworthy AI by providing a secure and transparent platform for storing and managing data and information related to AI systems and their decision-making processes. The decentralized nature of blockchains can help ensure that the data used to train and operate AI systems is secure, accurate, and tamper-proof. This, in turn, can help increase the level of trust in AI systems, as the data and processes underlying their decision-making can be audited and verified.

In addition, smart contracts, which are self-executing contracts with the terms of the agreement between buyer and seller being directly written into code, can be used to ensure that AI systems adhere to certain ethical and legal standards. For example, a smart contract could be used to automatically enforce data privacy policies, or to ensure that an AI system is not used for malicious purposes.

With the growing need for trustworthy AI, many different methods and frameworks have been proposed recently. Various methods focus on different stages of the AI lifecycle to make AI systems reliable and trustworthy. Some approaches focus on the design phase of the AI systems, which helps lay out the trustworthy requirements and expectations for AI systems. Some methods deal with the data collection, protection, and pre-processing phase, making data fair, diverse, and secure. Some approaches focus on the modeling phase of the AI system to provide explainability and interpretability of the system. Other methods work with the implementation and oversight phase of the AI system, which utilizes proper auditing and testing techniques to ensure accountability and reliability.

In this paper, we make three contributions. First, we present a comprehensive background, concepts, and need for a trustworthy AI system. Second, we review and organize the existing methods and guidelines that make AI systems trustworthy.

# Technical Background

## Trustworth AI Overview

Trustworthy AI refers to the use of Artificial Intelligence (AI) technologies and applications that are safe, secure, transparent, and ethically aligned with human values and rights. It encompasses a set of principles, processes, and methods aimed at ensuring that AI systems are developed and operated in a way that protects the well-being of individuals, society, and the environment, and that provides accountability and explainability. Key aspects of trustworthy AI include data privacy and security, bias and fairness, transparency and explainability, and ethical and legal compliance. The development of trustworthy AI requires collaboration across multiple domains including computer science, social sciences, ethics, and law.

## Preliminaries and Definitions

Trustworthy AI is not a monolithic concept but a polylithic one [XX]. Different terms in this field have several different interpretations. Therefore, it is imperative to define and explain these terms before we can use them. This section contains essential definitions of the terms related to the field of AI.

Artificial Intelligence: AI is a field that deals with making machines think. Legg and Hutter [XX] define AI as a process of imitating human behavior and decision-making capabilities. So, AI is a way to train machines to perform tasks that require intelligence.

Black-Box Problem: The black-box problem means that the system is opaque, and it is difficult to track the structure, internal working, and system implementation [XX]. AI systems are becoming more complicated, making them challenging to understand [XX]. This problem decreases the system’s trustworthiness as it is challenging to provide the reasoning and explanation for the output. system to different users with whom the system interacts

Explainable and Interpretable AI: Explainable and interpretable AI deals with developing explainable and interpretable models. Miller [XX] defines explainable AI as how an explanatory agent provides reasoning for their own or another agent’s decision making. Arrieta et al. [XX] describe explainable AI as a suite of algorithmic techniques that generate high-performance explainable models that humans can easily understand and trust. Researchers often use the terms explainability and interpretability interchangeably [XX]. So, in this article, we also use them interchangeably.

Reliability: Reliability of the system ensures that the system performs as intended—that is, within specified limits and without any failure, it produces the same outputs for the same inputs consistently [XX].

Fairness: Fairness of the system ensures that there is an absence of any discrimination or favoritism toward an individual or a group [XX] based on any inherent or acquired characteristics that are irrelevant in the context of decision making [XX].

Trust: Trust is a complex phenomenon [XX]. Different disciplines defined trust differently. Sociologists view trust as an attribute of human relationships [XX] and psychologists consider it as a cognitive attribute [XX], whereas economists think it is calculative [XX]. An agreement among these definitions is that trust has something to do with integrity and reliability. Philosophically, the National Institute of Standards and Technology [XX] defines trust as “the confidence one element has in another, that second element will behave as expected.”

Acceptance: Acceptance of an AI system is the user’s willingness to use the system in service encounters [XX].

Trustworthy AI: Trustworthy AI is a framework to ensure that a system is worthy of being trusted based on the evidence concerning its stated requirements. It makes sure that the users’ and stakeholders’ expectations are met in a verifiable way [XX].

## Need for Trustworth AI

These days, AI systems have achieved enough performance to be used widely in our society. These technologies are already transforming people’s lives [XX]. However, even though these AI systems have some utility, this does not imply that they are good enough and trustworthy. This informal attitude toward these systems is inappropriate when dealing with high-stakes applications where one wrong decision can lead to dangerous consequences. These systems can be brittle and unfair. Marcus and Davis [XX] provide an excellent example of facial recognition software that explains the need for trustworthy AI. If the facial recognition software is used for auto-tagging people in social media pictures, less reliable software is acceptable. Still, the same tool is unacceptable if the police want to use it to find suspects in surveillance photos. This example demonstrates how people adopt AI systems only when there are no life-critical consequences for them and society. To deliver AI benefits to high-stakes applications and increase AI systems’ adoption, we need an ethical framework to control and govern them.

## Framework of Trustworthy AI

Guidelines articulate a framework for achieving Trustworthy AI based on fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), and in relevant international human rights law. Below, we briefly touch upon Trustworthy AI’s three components

* Lawful AI

I systems do not operate in a lawless world. A number of legally binding rules at European, national and international level already apply or are relevant to the development, deployment and use of AI systems today.

The law provides both positive and negative obligations, which means that it should not only be interpreted with reference to what cannot be done, but also with reference to what should be done and what may be done. The law not only prohibits certain actions but also enables others. In this regard, it can be noted that the EU Charter contains articles on the ‘freedom to conduct a business’ and the ’freedom of the arts and sciences’, alongside articles addressing areas that we are more familiar with when looking to ensure AI’s trustworthiness, such as for instance data protection and non-discrimination.

* Ethical AI

Achieving Trustworthy AI requires not only compliance with the law, which is but one of its three components. Laws are not always up to speed with technological developments, can at times be out of step with ethical norms or may simply not be well suited to addressing certain issues. For AI systems to be trustworthy, they should hence also be ethical, ensuring alignment with ethical norms.

* Robust AI

Even if an ethical purpose is ensured, individuals and society must also be confident that AI systems will not cause any unintentional harm. Such systems should perform in a safe, secure and reliable manner, and safeguards should be foreseen to prevent any unintended adverse impacts. It is therefore important to ensure that AI systems are robust. Ethical and robust AI are hence closely intertwined and complement each other.

## Implementation and realisation of trustworthy AI

Different groups of stakeholders have different roles to play in ensuring that the requirements are met:

* Developers should implement and apply the requirements to design and development processes;
* Deployers should ensure that the systems they use and the products and services they offer meet the requirements;
* End-users and the broader society should be informed about these requirements and able to request that they are upheld.

## Requirement to make Trustworthy AI

A trustworthy approach is key to enabling “responsible competitiveness”, by providing the foundation upon which all those affected by AI systems can trust that their design, development and use are lawful, ethical and robust. These Guidelines are intended to foster responsible and sustainable AI innovation in Europe.

They seek to make ethics a core pillar for developing a unique approach to AI, one that aims to benefit, empower and protect both individual human flourishing and the common good of society. We believe that this will enable Europe to position itself as a global leader in cutting-edge *AI* worthy of our individual and collective trust. Only by ensuring trustworthiness will European individuals fully reap AI systems’ benefits, secure in the knowledge that measures are in place to safeguard against their potential risks.

The below list of requirements is non-exhaustive. It includes systemic, individual and societal aspects:

* Human agency and oversight: AI systems should support human autonomy and decision-making, with human oversight and intervention as integral elements.
* Technical robustness and safety: AI systems have to be resilient, reliable and secure, developed with a focus on preventing and minimizing unintended harm.
* Privacy and data governance: AI systems should provide adequate governance in terms of privacy and data protection, and quality, integrity, and access to data.
* Transparency: the data, system, and business models of AI systems should be transparent and explainable for stakeholders. Humans need to be informed when they interact with an AI system, and apprised of its capabilities and limitations.
* Diversity, non-discrimination and fairness: AI systems need to avoid unfair bias and provide for accessibility and universal design. Stakeholders who may be affected by an AI system should be considered and involved.
* Societal and environmental well-being: AI systems should be sustainable, environmentally friendly, considering broader society and other sentient beings. The impact on institutions and democracy also needs to be taken into account.
* Accountability: Mechanisms for ensuring responsibility, accountability, and potential redress for AI systems and their outcomes should be put into place. Negative impacts should be identified, assessed, documented, and minimized.

## Overview of Blockchain

Blockchain is a decentralized digital ledger that is used to record transactions across multiple computers in a secure and transparent manner. It is essentially a chain of blocks, where each block contains a record of multiple transactions, and each block is linked to the previous block through cryptography. The decentralization of the blockchain ledger means that there is no single point of control or failure, which makes it more secure and resilient compared to traditional centralized systems.

Blockchain technology was originally developed as the underlying technology for the cryptocurrency, Bitcoin, but it has since been adopted for a wide range of other use cases, including supply chain management, digital identity, and voting systems.

One of the key benefits of blockchain technology is its ability to provide a secure and transparent record of transactions without the need for intermediaries. This allows for faster, cheaper, and more secure transactions, as well as increased transparency and accountability.

There are several types of blockchain technology, including public blockchains, private blockchains, and consortium blockchains. Public blockchains are open and accessible to anyone, while private blockchains are only accessible to a select group of participants. Consortium blockchains are a hybrid of public and private blockchains, where a group of organizations come together to control the blockchain.

Overall, blockchain technology has the potential to transform a wide range of industries and has already begun to disrupt traditional business models. However, there are also challenges associated with the adoption of blockchain technology, including scalability, interoperability, and regulatory issues.

## Features of blockchain

Blockchain technology has several key features that make it unique and valuable for various use cases:

* Decentralization: Blockchain is decentralized, meaning that there is no central authority or single point of control. This makes the system more secure and resilient, as there is no single point of failure.
* Immutable records: Once a transaction is recorded on the blockchain, it cannot be altered. This creates a permanent and unalterable record of transactions, which can increase transparency and accountability.
* Secure transactions: Blockchain uses cryptography to secure transactions and prevent unauthorized access. This helps to ensure the privacy and security of transactions.
* Transparency: Transactions recorded on the blockchain are visible to all participants, making the system transparent. This can increase trust in the system and make it easier to detect fraudulent or malicious activity.
* Interoperability: Blockchain technology is interoperable, meaning that it can communicate with other systems. This allows for the exchange of data and value between different blockchain systems.
* Smart contracts: Blockchains support the use of smart contracts, which are self-executing code snippets that enforce the terms of a contract. This allows for automatic and secure execution of agreements, without the need for intermediaries.
* Consensus mechanism: Blockchains use a consensus mechanism to ensure that all participants agree on the state of the ledger. This helps to prevent disputes and ensure the integrity of the system.

These features make blockchain technology useful for a wide range of use cases, including supply chain management, digital identity, voting systems, and financial services. The decentralized and secure nature of blockchain technology makes it well-suited for applications where trust is critical.

## How does blockchain enable trustworthy AI

Blockchain enables trustworthy AI by providing a decentralized and tamper-proof ledger for recording and storing data, as well as executing smart contracts. This ensures that the data used to train AI models is secure, transparent, and immutable, preventing malicious actors from manipulating the data and compromising the accuracy of the models. Additionally, the use of smart contracts on the blockchain can enforce specific rules and conditions for how AI models can be used and deployed, providing an additional layer of accountability and trust.

The implementation of blockchain technology builds trust between anonymous users and enables them to perform transactions without any third-party intermediary [XX].

The future scope of blockchain technology can be incorporated into big data and standardization. The integration of blockchain technology with big data helps to

keep confidential data secrets and copyright forms in the secured cryptographic layer.

The increasing computational power and proliferation of big data are now empowering Artificial Intelligence (AI) to achieve massive adoption and applicability in many fields. The lack of explanation when it comes to the decisions made by today's AI algorithms is a major drawback in critical decision-making systems. Explainable AI (XAI) is a new trend of AI algorithms that provide explanations of their AI decisions [12]

Technical trustworthiness as a measure of acceptance of decisions by DApp users have requirements of consensus, economic models and incentives for honesty, explainability, and robustness of predictors. In addition, many more infrastructure requirements are needed such as security, privacy, reliability, usability, dependability, performance, and governance. The emerging blockchain technology seems the most adequate, if not the only one, to fulfil these requirements. Still, many challenges must be tackled, the most important ones are minimizing human in the loop for validating explanations and real timeliness for certain applications.[XX]

When Blockchain is combined with trustworthy AI, it has the potential to bring new levels of security and transparency to various industries. By incorporating AI algorithms into the blockchain, data can be analysed and processed in real-time, providing valuable insights and reducing the risk of fraud. Furthermore, the decentralized nature of blockchain ensures that the data is not controlled by a single entity, reducing the risk of data manipulation and increasing the overall trust in the system. In this way, blockchain and trustworthy AI can work together to create a secure and transparent ecosystem that benefits both individuals and organizations.

# Related work

There are several frameworks available that aim to enforce the principles of trustworthy AI. Some examples include:

* The European Commission's AI ethics framework, which provides guidance for ethical AI development, deployment and use, based on the seven principles of trustworthy AI.
* The MIT-IBM Watson AI Lab's AI Accountability framework, which provides a set of principles and practices for ensuring accountability in AI systems.
* The Partnership on AI's principles for responsible AI, which outline best practices and guidelines for responsible AI design, development, deployment and use.
* The Algorithm Accountability Act, a proposed piece of legislation in the US aimed at ensuring that AI algorithms are transparent, explainable and fair.

The IEEE Global Initiative for Ethical Considerations in AI and Autonomous Systems, which provides guidelines and standards for ethical AI design and use.

These frameworks can serve as useful tools for organizations and individuals looking to develop and implement trustworthy AI systems. However, it is important to note that the actual enforcement of these principles depends on the regulations and laws in place in different jurisdictions.

There are several frameworks available to help implement the principles of trustworthy AI. Some examples include:

* The FAIR (Findable, Accessible, Interoperable, Reusable) Guiding Principles for scientific data management and stewardship, which can be applied to AI systems to ensure the responsible management of data and algorithms.
* The AI HLEG's Recommendations on a Common European Approach to AI, which provide guidance on the development and deployment of AI systems, including the implementation of ethical principles.
* The International Organization for Standardization (ISO) standards on AI, which provide guidelines for the development and use of AI in a responsible and ethical manner.
* The AI Now Institute's AI Policy and Practice Playbook, which provides a practical guide for organizations and policymakers looking to implement ethical and responsible AI practices.
* The AI Alignment Forum's Long-Term AI Strategy, which outlines a strategy for the responsible development and use of AI, based on principles of transparency, accountability and value alignment.

These frameworks can help organizations and individuals ensure that AI systems are designed and used in a way that is consistent with the principles of trustworthy AI. However, it is important to note that the implementation of these frameworks will depend on the specific circumstances and context of each individual case.

There has been substantial work done to develop frameworks to support the technical implementation of Trustworthy AI, but these frameworks have been very use case specific and do not always encompass all seven principles.

Nassar et. al developed a framework based on the premise that the critical decisions in complex AI systems must be subject to a consensus among distributed AI and XAI agents or predictors hosted by trusted oracles with the assumption that the majority of these agents are honest. Trustworthy AI requirements for resilience to biases and adversarial attacks can be fulfilled to a large extent by blockchain technologies. They presented several use cases to show how blockchain SCs combined with decentralized storage can be leveraged to achieve a trustworthy XAI.

Abhishek et. al proposed a framework for trustworthy AI systems in the age of pervasive data collection and computing. This framework provides a data-centric level of abstractions for ethical questions posed in the AI and Data Science context. This framework focused on the ethics of data and ethics of algorithms, and more specifically, the aspects that can be integrated directly within the design and development of AI systems.

Bo. Li et. al tried to unify currently available but fragmented approaches toward trustworthy AI considering the entire lifecycle of AI systems, ranging from data acquisition to model development, to system development and deployment, finally to continuous monitoring and governance. They offered concrete action items for practitioners and societal stakeholders (e.g., researchers, engineers, and regulators) to improve AI trustworthiness. Finally, they identifed key opportunities and challenges for the future development of trustworthy AI systems, where we identify the need for a paradigm shift toward comprehensively trustworthy AI systems.

The world already moves towards the 6G era. AI/ML mechanisms will become structural components of the system and operate in a native manner. As the systems get more complex and intelligent, mechanisms for ensuring trust in those operations become critical.

Barmpounakis et. al justified the need for a framework for protecting the input of the AI mechanisms, for achieving an explainable operation, and for guaranteeing proper outputs in the areas of Mobile network management.

Wajid et. al introduced a digital manufacturing platform architecture that extends Industry 4.0 paradigms to enable AI-based decision support with the necessary trustworthiness and human-centricity elements primed for Industry 5.0. The proposed architecture helps realize the balancing act of getting the perceived benefits from AI-centric digitalization while preserving the role of humans in key decision-making activities. They suggested the use of technologies targeting concepts like human-centricity, sustainability and resilience to move from Industry 4.0 to 5.0

Ahuja et. al suggested that Trustworthiness is a central requirement for the acceptance and success of human-centred artificial intelligence (AI). To deem an AI system as trustworthy, it is crucial to assess its behaviour and characteristics against a gold standard of Trustworthy AI, consisting of guidelines, requirements, or only expectations

The realization of trustworthy AI systems is one of the big challenges for the success of ethical and human-centered AI. This has been acknowledged by both politics [XX] and academia. For the implementation of trustworthiness principles, we argue for the adoption

of methods and technologies from software engineering. Software engineering has a long-standing tradition on the principled construction of complex systems and has already much of the fundamental work available.

# conclusion

IN PROGRESS

- focus on limitations on what features are covered - underline the gap - emphasize - no generic framework - open the door to your contribution

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**