Göksal, Şaban-İbrahim; Solarte-Vasquez, Maria Claudia

Article

The blockchain-based trustworthy artificial intelligence supported by stakeholders-in-the-loop model

Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration

Provided in Cooperation with:
Faculty of Economics and Administration, University of Pardubice

Suggested Citation: Göksal, Şaban-İbrahim; Solarte-Vasquez, Maria Claudia (2024) : The blockchain-based trustworthy artificial intelligence supported by stakeholders-in-the-loop model, Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration, ISSN 1804-8048, University of Pardubice, Pardubice, Vol. 32, Iss. 2, pp. 1-19, https://doi.org/10.46585/sp32022083

This Version is available at:
https://hdl.handle.net/10419/334841

Article

SciPap
scientific papers

Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration
2024, 32(2), 2083.

DOI: 10.46585/sp32022083
editorial.upce.cz/SciPap

The Blockchain-Based Trustworthy Artificial Intelligence Supported by Stakeholders-In-The-Loop Model

Şaban İbrahim Göksal

Tallinn University of Technology, Department of Law, School of Business, Estonia

Maria Claudia Solarte Vasquez

Tallinn University of Technology, Department of Law, School of Business, Estonia

Abstract

This paper introduces the Blockchain-Based Trustworthy Artificial Intelligence (AI) Supported by Stakeholders-In-The-Loop Model (BCTrustAI.SL) that incorporates sociotechnical components to ensure legal compliance and conformity with the emerging trustworthiness standards for AI systems. BCTrustAI.SL combines features of the Blockchain Framework for Trustworthy AI (BF.TAI), the BlockIoTIntelligence architectural model and the Society-In-The-Loop (SITL) framework, to solve the intangibility problem of the foundational normative notions of robustness, ethicality and lawfulness, and human-centredness, and help them become applicable in practice. The work contributes to the specification and operationalisation of trustworthiness as an AI attribute, highlighting the importance of a precise understanding of its institutional foundations, high order principles and concrete key requirements (data protection, data governance, technical robustness and safety, transparency, accountability, diversity and non-discrimination, and human agency and oversight) deriving from them. On a practical/technical level, BCTrustAI.SL showcases the strengths of combining Blockchain (BC) and AI to address their individual limitations, laying the groundwork for future advancements and practical applications.

Keywords

Trustworthy AI, Robustness, Ethicality and Lawfulness, Human-centredness, Blockchain Framework for Trustworthy AI, BlockIoTIntelligence Architectural Model, Society-In-The-Loop Framework

JEL Classification

K39, M15, M48, O31, O33, O38

Introduction

AI technology faced a stagnation period coined by Minsky and Schank in 1984 as the 'AI winter' (Umbrello, 2021, pp. 7-8), but has since experienced phenomenal growth. Major developments include reinforcement learning (Sutton et al., 2018, pp. 1-13), deep neural network architectures, and large language models (LLMs) such as ChatGPT and similar agents (Russell & Norvig, 2010, pp. 34-59& 64-108; Ram et al., 2018; Bubeck et al., 2023). The ability to process large datasets drives the continued growth of AI, making it an indispensable part of our daily lives and revolutionising all industries. While the capabilities of machines have advanced from being merely computational and predictive, to becoming relational and proactive (Meurisch et al., 2020; Das et al., 2023), mistakes such as mislabelling and displaying biased outputs still occur, profoundly affecting individuals and the exercise and enjoyment of their fundamental rights. Instances that made it to the headlines are the outrageous labelling results of Google Vision Cloud, that suggested racial bias (Kayser-Bril, 2020), concerning errors or blindness when it came to recognising certain animals (Simonite, 2018; Yapo & Weiss, 2018, pp. 5366-5368; Krishnan, 2020, pp.496-497). In addition, machines are not hard to fool. Athalye et al. (2018) showed the flaws of AI-based vision recognition, which mistook 3D-printed turtles as rifles, and classified baseballs as 'espresso.' Besides, the potential bias and/or bias amplification problem (Pasquale, 2015, pp. 38-58; O'Neil, 2016, pp. 105-122& 141-160; Caliskan et al., 2017; Buolamwini & Gebru, 2018), other limitations that are found to erode the stakeholders' trust are opacity of decision-making processes (Rai, 2020; von Eschenbach, 2021), and lack of common-sense reasoning (Marcus, 2018). In sum, while AI outperforms humans in specific tasks; on the whole it merely simulates human cognition and perception (Marcus, 2018, pp. 6-7).

Public and private sector entities around the world have made important regulatory efforts to address these

Corresponding author:
Şaban İbrahim, Göksal, Akadeemia tee 3, 12616 Tallinn, Estonia
Email: saban.goksal@taltech.ee

limitations and ensure that the results of the development of AI are trustworthy (Antonov & Kerikmäe, 2020; Smuha, 2021). However, these efforts are not harmonized and abstract , creating challenges for practical implementation Examples of the former are the updated National AI, Research and Development Strategic Plan 2023 Update , the Coordinated Plan on Artificial Intelligence 2021 Review, and the Next Generation Artificial Intelligence Development Plan, published in the United States (US) by the Select Committee on AI of the National Science and Technology Council (2023), in the European Union (EU) by the EU Commission (EC) (2021), and in China by the Department of International Cooperation Ministry of Science and Technology (MOST) (2017), respectively. Their focus is placed on the ethical, legal, and societal implications of AI within R&D initiatives. In addition, the US Second Draft of the AI Risk Management Framework (National Institute of Standards and Technology, US Department of Commerce, 2022) and the EU AI Act (Regulation (EU) 2024/1689) established some standards and assessment tools applicable to the adoption and use of AI (Göksal et al., in press).

The US, the EU and China, have also published various guidelines (Trustworthy AI (TAI) Play Book (United States Department of Health & Human Services (US DHHS), 2021), EU Ethic Guidelines for Trustworthy AI (Independent High-Level Expert Group on Artificial Intelligence (AI HLEG), 2019)), a roadmap and other administrative orders (e.g. Memorandum-Guide for Regulation of Artificial Intelligence Applications (Executive Office of the President, Office of Management and Budget [OMB], 2020), Executive Order for Safe, Secure, and Trustworthy Artificial Intelligence (The White House, 2023), Responsible Artificial Intelligence Strategy and Implementation Pathway (United States Department of Defence (US DoD), 2022), etc.), and white papers (the White Paper on Artificial Intelligence: An European Approach to Excellence and Trust (European Commission, 2020), and the White Paper on Trustworthy Artificial Intelligence (China Academy of Information and Communications Technology [CAICT] & JD Explore Academy, 2021)) on the development and adoption of trustworthy AI.

These frameworks substantiate the first of the two stages of this work solving a research problem dimension on the specification of AI trustworthiness, potentially producing a more inclusive concept than the AI HLEG or any other on its own. In here, this concept is seen as a feature grounded on robustness, ethicality and lawfulness, and human-centredness; consistent with 3 deriving principles (harm prevention, fairness, and human autonomy); and, meeting 7 key requirements that correspond to the AI HLEG formulation: data protection, data governance, technical robustness and safety, transparency, accountability, diversity and non-discrimination, and human agency and oversight (AI HLEG, 2019, pp. 14-20). This clarification addresses ongoing regulatory ambiguities, suggesting a simpler but cohesive understanding for the operationalisation of terms that may easily spread beyond the European borders. The result of this specification is a comprehensive construct that facilitates rule determination and compliance in the elaboration of technical models and non-technical AI governance strategies as the second stage demonstrates. It relies on public initiatives from the US, the EU and China because of their leadership and the unique insights each offers for the advancement of the regulatory treatment of AI. The US proposes a market-oriented viewpoint (Hine & Floridi, 2022, pp. 4-7), the EU's is characterised by its human-centric emphasis (European Commission, 2020), and China's is pivotal on incentivising technological innovation (Hine & Floridi, 2022, pp. 8-12).

This and the second stage are self-containing in terms of problem dimension it solves, its corresponding methodological approach, results and contributions, but together they follow a constructivist logic where a partial institutional analysis (Joamets & Vasquez, 2020, pp. 112 & 115-122) as outlined above precedes, founding the architectural design of an AI training and data analysis model. Both stages rely on a standard literature review integrated in the institutional analysis discussion and the revision of the state of the art of the second stage, correspondingly.

Several technical and sociotechnical proposals for trustworthy AI are compatible with the regulatory landscape outlined, but three are representative for having gained wide acceptance are selected to show that the expected alignment with the principles and key requirements of trustworthy AI is not consolidated. Two use BC technology, as the means to enhance transparency, data governance, and technical robustness and safety (Nassar et al., 2020; Singh et al., 2020; Zhang et al., 2023) without letting BC diminish their capacity to incorporate human agency and oversight operations (Nassar et al., 2020, pp. 6-7). The BF.TAI (Nassar et al., 2020), which is set on achieving trustworthiness by meeting the key requirement transparency with the help of AI and Explainable AI (XAI) predictors, and the BlockIoTIntelligence architectural model (Singh et al., 2020) that assembles a sophisticated architecture to increase data governance, and technical robustness and safety in data collection and processing for IoT (Internet of Things) networks. The third is the SITL framework, that addresses human-centredness deficits facilitating human agency and oversight, and diversity and non-discrimination, by encouraging 'society' to improve the outputs of AI, thereby participating in decision-making (Rahwan, 2018).

The second research problem dimension is represented by these proposals which fall short in the operationalisation of the features of trustworthy AI for legal compliance and conformity with principles. Issues appear to loom large with the first two because they neglect the post-deployment trust problems (related to human agency and oversight) whereas regarding the SITL framework, questions arise about how to guarantee control and diminish bias (Martínez Ramil, 2021, p.5; Cunningham & Delany, 2021) as well as about underfitting (Cunningham & Delany, 2021) if the model supposes one type of intervention, prior the retraining, and assumes that the formation of an algorithmic

social contract is possible, based on an unchecked and loosely described 'determination' process (Rahwan, 2018, pp. 9-10). None is particularly concerned with accessibility and communication between the front-end users and the systems as instrumental to human agency and oversight or considers transparency criteria as the essential element it is for procedural fairness in contemporary regulatory systems (Lee et al., 2019).

Alternatively, the BCTrustAI.SL and blueprint of a trustworthiness-by-design model this paper introduce, feature transparency and autonomy as the BF.TAI, and technical robustness and safety like the BlockIoTIntelligence architectural model, but making it extensive to all data collection, storage, and management cycles and layers of the system. In addition, acknowledging the need to increase human-centredness, according to the key requirement of human agency and oversight, BCTrustAI.SL adjusts the SITL's approach to mitigate some of the trust issues that are likely to arise post-deployment, collecting multiple evaluation data in a feedback loop from focused groups, without subscribing to a generic 'collective agreement'. The optimised system enables a diversity of stakeholders to be 'in the loop' and is open to various opportunities for user centricity adjustments, showcasing an adaptable design. This means that the model augments accessibility and communication between the users and the systems as conditions of fairness and transparency, even in the narrow sense of UX/UXI attribution factors, screening legally relevant interactions. The combination and increased capacities of BCTrustAI.SL offers a solution to help realise essential regulatory goals and makes sense of expressions such as human-centredness, human-in-the-loop (HITL) (AI HLEG, 2019, pp. 15-16; Wu et al., 2022, p.2), human-on-the-loop (HOTL) (AI HLEG, 2019, pp. 15-16; Li et al., 2020), and Human in Command (HIC) (AI HLEG, 2019, pp. 15-16). In sum, BCTrustAI.SL permits humans to participate in various stages in the AI development process to exercise an adequate and truly informed oversight on AI from development and deployment to use.

The following sections restates the methodology and explains the theoretical and empirical work leading to the design of BCTrustAI.SL. The third consists of the delineation of the conceptual and formal institutional background resulting in the normative synthesis of trustworthy AI. The fourth unpacks and discusses the 'trustworthy AI' models representing the state of the art in the related literature, detailing selected features that inform the BCTrustAI.SL design. The fifth section delivers the blueprint and working flow of the model. The sixth reflects on its fit to the regulatory purposes. The last restates and outlines theoretical and practical contributions, limitations, and potential for future research.

Methodology

The research design is constructivist (Wadsworth, 2003) and combines empirical and theoretical methods to fulfil the purpose and aims of the study. Conceived to follow a two-stage, multimodal strategy, as shown in **Fig. 1**, it begins with a partial institutional analysis (Joamets & Vasquez, 2020, pp. 112 & 115-122) which relies on public normative sources and definitions from the literature to refine and expand the concept of trustworthy AI. The resulting normative construct grounds the operationalization effort guiding the subsequent design of an AI training and data analysis model. The second stage proceeds to describe and discuss three widely accepted system architectures, to point out that in terms of compliance and conformity with the emerging normative framework on AI, the state of the art is still limited. The methods in use are a standard literature review and deep content analysis. To end, this stage produces BCTrustAI.SL, an advanced sociotechnical model that surpasses earlier proposals, based on an AI and ML model design process. The model is unique in that it permits a wider operationalization of trustworthiness features, as they are being enshrined in the institutional discourse.

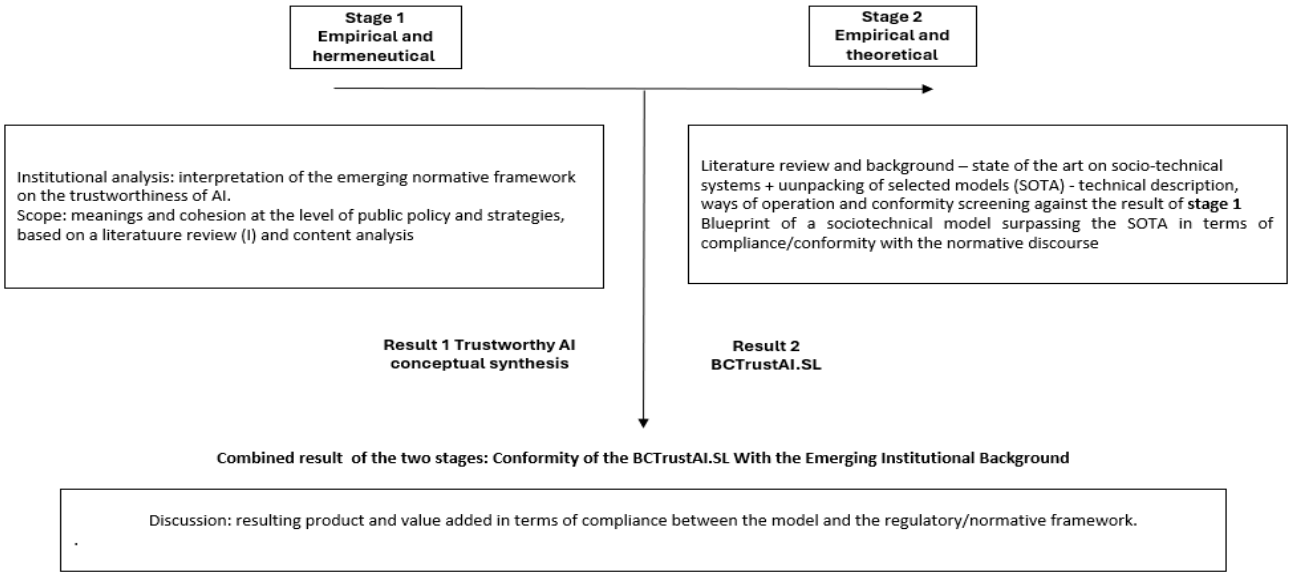

Fig. 1. Research design.

Theoretical Background. Conceptual and Institutional Review

This section covers the non-technical components of the trustworthy AI concept as they have been emerging in the literature and in the international and supranational institutional environments. 'Institutional environment' refers to formal frameworks encompassing laws, policies, and administrative rules governing the development, implementation, and use of AI technologies. Narrowly, it reflects a set of trustworthy AI properties (catalogued as characteristics, foundations, principles, and key requirements), representing major regulatory trends and proposals dedicated to the treatment of AI and future technologies globally.

The institutionalisation of 'Trustworthiness' as a Compliance Attribute of AI

A strong understanding of the process leading to the institutionalisation of trustworthiness in this context requires the revision of perspectives from different regions and their synthesis. In addition, to prime an institutional analysis when phenomena under study are inadequately defined, regulatory developments are incomplete, and compliance assessment methodologies are not harmonised (Joamets & Vasquez, 2020, pp. 112 & 115-122), calls for a start on terminology delimitation. Efforts have been made to render trustworthy AI and neighbouring concepts more tangible, resulting on practical upgrades to respond to some of the trust questions associated with these technologies (Jobin et al., 2019, pp. 13-18; Hine & Floridi, 2022, pp. 689-690; Hohma & Lütge, 2023, pp. 904-910), but achieving a cohesive understanding across fields, according to a judicious survey of the literature remains a challenge.

With this in view, the analysis begins examining the meaning of trustworthiness in public sector-driven normative documents from the US, the EU, and China. These perspectives are of significance because they represent influential regulatory trends (European Commission, 2020; Hine & Floridi, 2022, pp. 692-700) deriving from well-conceived and yet diverse stances. The first is market oriented, the second is centred on human rights, and the third follows a strategic innovation system pathway.

The conceptual specification this paper recommends, perceives the nuanced nature of trustworthiness, considering it an attribute, that indicates the AI systems' legality, ethicality and legitimacy. This approach helps reduce uncertainty and prevents interpretative incongruences (Hagendorff, 2020, pp. 103-106; Jobin et al., 2019, pp. 13-18). Fig. 2 illustrates the regulatory layers, components (characteristics, principles, and key requirements) and interconnectedness demarcating trustworthiness that were identified in the US responsible AI strategy, the EU Ethics Guideline for Trustworthy AI and the White Paper on Trustworthy AI from China.

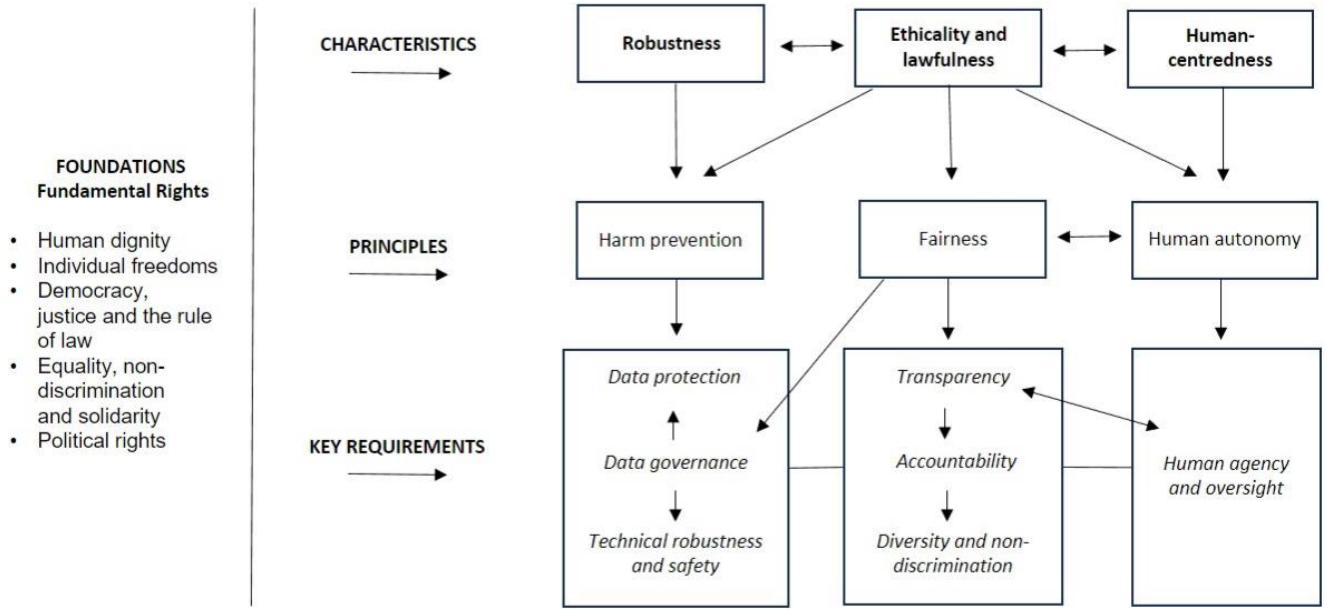

Fig. 2. Regulatory components of trustworthiness.
Source: charted by the authors

The first layer groups the most general features that should characterise AI trustworthiness: robustness, ethicality and lawfulness, and human-centredness. These serve as the broadest criteria. The Ethics Guideline released by the AI HLEG and the Chinese White Paper describe these aspects as 'dimensions' and 'components' of trustworthy AI. In contrast, the US AI strategy has referred to three regulatory spheres concerned with trustworthiness: the legal and the ethical frameworks, and policy guidelines.

The second layer consists of the three foundational principles: harm prevention, fairness, and human autonomy, which are connected to the five fundamental rights that the EU Ethics Guideline (AI HLEG, 2019 pp. 10-11)

highlights to convey the idea that technology development should inherently uphold democratic principles. In contrast to other documents that speak of principles as requirements themselves, their primary role in here, lies in drawing the scope of allegiance intended by establishing the key requirements below.

The third layer lists the seven key requirements intended to operationalise trustworthiness principles and to keep AI within ethical boundaries during AI development processes: data protection, data governance, technical robustness and safety, transparency, accountability, diversity and non-discrimination, and human agency and oversight. Environmental well-being is not considered as a key requirement due to its highly abstract nature and the complex trade-off between the energy-intensive operations of AI and BC technologies and the potential environmental benefits they can offer (M. Bublitz et al., 2019; Taghikhah et al., 2022). This layer continues to draw from the EU guidelines, because in the US and China, these are listed as principles and characteristics.

Characteristics, Principles and Key Requirements

Robustness, as a characteristic of trustworthiness, refers to resiliency, reliability, and performance capabilities of AI systems under varying conditions and potential adversities (AI HLEG, 2019, p. 7). The emphasis on robustness is a security matter, driven by the need to ensure that models behave predictably and can withstand changes in their environment or input data (Goodfellow et al., 2015, pp. 2-7). This includes protection against adversarial attacks that could interfere with AI systems' outputs (Szegedy et al., 2014). Robustness is attained when modelling strategies, fulfilling the key requirements data protection, data governance and technical robustness and safety, are in conformity to the harm prevention principle. The key requirements were adopted in the US TAI Play Book (US DHHS, 2021, pp. 21-23) as 'safety and security', 'privacy' and 'robustness' and reliability', in the Responsible AI Strategy and Implementation Pathway (US DoD, 2022, pp. 5-7), as 'governable', and in the Chinese White Paper on Trustworthy AI (CAICT & JD Explore Academy, 2021, pp. 6-7) as 'reliability' and 'data protection'.

Data governance plays a chief safeguarding role in view of the vulnerability of data to adversarial attacks (Janssen et al., 2020; Atoum & Keshta, 2021), Yang et al. (2019 pp. 2-3 & 10-11) defines it from a cybersecurity perspective as the systematic management of data throughout its lifecycle to ensure quality, security, accessibility, and compliance within particular regulatory frameworks. It pertains to the establishment and enforcement of structured protocols, as well as the assignation of concrete responsibilities to overseeing authorities for the effective control of the data and the algorithms (Janssen et al., 2020, pp. 1-3). Data governance including data protection, entails measures leading to technical robustness and safety. Adherence to these key requirements is anticipated to make AI systems resilient against external interference and manipulation and equipped with recovery capacities when faced with such challenges. Safety measures such as authentication processes and encryption, generally guard against unauthorised access or alteration of the data (AI HLEG, 2019, pp. 16-17). According to Meurisch & Mühlhäuser (2021, p.6), data protection refers to technical measures and approaches that secure personal data against unauthorised access and breaches (AI HLEG, 2019, p. 17). Data governance is instrumental to the fairness principle under ethics and lawfulness because it is linked to legal provisions including the rights and duties enshrined in the General Data Protection Regulation of the EU (GDPR) (Reg 2016/679). At the same time, it is a precondition of transparency and accountability that are not only key requirements but also features that well-known theoretical classifications assign to reliable AI systems (Dignum, 2019, pp. 52-62).

Ethicality and lawfulness work as an overarching characteristic of the combined frameworks since these speak of an orderly and interrelated value driven catalogue of standards highlighted in laws and/or official policy documents. They raise the bar of trustworthiness for AI adding criteria beyond the current legal standards. The aim is to steer AI development with a dynamic approach, rather than adhering to rigid, traditional regulations alone. To build public trust in technology, and boost adoption, AI models must protect individual rights (Russell et al., 2016; Floridi et al., 2018; Dignum, 2018) and make achievable several fairness categories represented by the three key requirements deriving from this principle: transparency, accountability, and diversity and non-discrimination. Fairness and human autonomy, which amount to individual freedom cannot be interpreted separately.

On the first key requirement of this set, whereas 'traceability' is a requirement in the US TAI Play Book (US DHHS, 2021, p.19), and 'transparency' and 'explainability' in the Responsible AI Strategy and Implementation Pathway (US DoD, 2022, pp. 5-7), and the Chinese White Paper on Trustworthy AI (CAICT & JD Explore Academy, 2021, p.6-7). Accountability is adopted as 'responsibility' and 'accountability' in the US TAI Play Book (US DHHS, 2021, p. 20), as 'responsibility' in the Responsible AI Strategy and Implementation Pathway (US DoD, 2022, pp. 5-7), and as 'clear responsibility' in the Chinese White Paper on Trustworthy AI as well (CAICT & JD Explore Academy, 2021, pp. 6-7). Diversity and non-discrimination is similar to a requirement called 'fair and impartial' in the US TAI Play Book (US DHHS, 2021, p. 18), to 'equitable' in the Responsible AI Strategy and Implementation Pathway (US DoD, 2022, pp. 5-7), and to 'diversity and tolerance' in the Chinese White Paper on Trustworthy AI (CAICT & JD Explore Academy, 2021, pp. 6-7).

Conceptually, transparency denotes clarity and openness, which in the context of AI systems are vital for trust. It demands insight into how the system functions, the data and algorithms it uses, and the processes justifying the machine's outputs. This quality is determined by a bundle of factors that measure the accessibility of the stakeholders to the system's operations and the data governance mechanisms it employs (Dignum, 2019, p.54).

The related terms that the literature and policy documents mention are ‘traceability’, ‘explainability’, and ‘communication’. ‘Traceability’ is keeping AI input data and processes’ records while ‘explainability’ is defined as one aspect of transparency about the system's openness in respect to processing activities and outputs; it indicates the system’s ability to present its decisions or actions in a manner interpretable by humans to counterbalance the ‘Blackbox’ phenomenon (von Eschenbach, 2021). The AI HLEG (2019, p. 18) The AI HLEG [25, p. 18] states that ‘communication’ is a feature that makes AI systems recognisable, and their limitations and capabilities clearly conveyed. In sum, transparency becomes a precondition of the operationalisation of human-centredness in terms of human agency and oversight.

The literature suggests that there is a strong link between transparency and accountability (Ananny & Crawford, 2018, pp. 976 & 982-983; Felzmann et al., 2020, p. 3338), and so does the Ethics Guideline for Trustworthy AI (AI HLEG, 2019 pp. 19-20), but we lack a consolidated definition of accountability as well as the methodology and parameters to assess it. Novelli et al. (2023, p.2) say that accountability is an obligation of an agent to justify its ‘conduct’ to whomever hold the authority to supervise, question, and assess it, especially in the case of delegated tasks. If to add that the disclosure should demonstrate that the system's design and development has considered the broad implications it may have on society, stakeholders, and human values (Dignum, 2019, pp. 53-56), their definition works for this re-conceptualisation on trustworthiness.

Diversity and non-discrimination mean the inclusive representation of various groups in AI development and applications, to prevent and reduce bias, like benefiting or harming any group or person for reasons of on race, gender, age, or other protected categories (Martínez Ramil, 2021, pp. 3-8; Cachat-Rosset & Klarsfeld, 2023, pp. 2-6).

There is a direct connection between data governance and these key requirements. Diverse and non-discriminatory AI systems, for instance, can only result from following a robust data governance strategy, and while upholding foundational human rights, conformity with the requirement would be supporting the development of human centric AI by design (Dignum, 2019, pp. 62-67). Transparency in AI on the other hand, is vital for informed decision-making, and accountability, enabling effective human agency and oversight (Lehner et al., 2022, pp. 120-125). Conversely, human involvement ensures AI systems are explainable, ethically aligned, and continuously improved, reinforcing the need for transparency. Essentially, transparency and human agency and oversight are interdependent, each amplifying the importance of the other in the realm of AI (von Eschenbach, 2021, pp. 1615-1618; Mosqueira-Rey et al., 2023, pp. 3032-3038).

Human-centredness is the last characteristic of trustworthy AI, and one that newly recognises the essential role of human-AI interaction and cooperation experiences in building a digital and automated ecosystem of trust. Broadly, human-centredness implies that AI systems enhance and extend human capabilities and the general good. Hence, the multiplicity of neighbouring expressions finding way in the related literature such as user centricity (Solarte-Vásquez & Nyman-Metcalf, 2017, p. 22), society-centredness (Ishida, 2004; pp. 16-17), stakeholders’ centricity (Taylor et al., 2023, pp. 3-5), and so on. Neutral to those discussions, the institutionalisation of trustworthiness is generally limited to components mentioned in the chart above, circumscribing human-centredness to a notion of responsible (ethical and lawful), and ‘human-compatible’ AI. The EU framework ascribes to this characteristic the principle of human autonomy, which is respected if the humans are in actual control of the systems, can exercise a competent oversight and evaluation of their design, development, implementation and operations’ processes (AI HLEG, 2019, pp. 15-16).

Human agency is a mix of empowerment, skills and transactional capacities of human users when interacting with AI systems whereas Human oversight pertains to the supervisory role that they play in monitoring and intervening in AI operations. Oversight is performed keeping HIC for overarching control over AI processes, HITL for direct intervention in AI decisions, and/or HOTL for supervisory roles with potential intervention (AI HLEG, 2019, p.15; Fanni et al., 2023, pp. 3-5). Rahwan (2018, pp. 6-7) argues that HITL, HOTL, or HIC are not sufficient for social dialogue, and that SITL is as critical, as does the text of the EU policy document which states that the risk assessment should be conducted through social dialogue. Along with Rahwan (2018, pp. 7-10) and the AI HLEG, Zicari et al. endorse large stakeholders' participation with their Z- Inspection assessment process (Zicari et al., 2021, pp. 84-89).

Respect for fundamental rights was strongly emphasised when defining human-centredness by the AI HLEG and it has thus come to rank high in the evaluation of AI systems and the core characterisation of the EU institutional framework on AI and future technologies. It has to do with devising continuous risk assessment techniques towards minimising potential harm, which conveys the need for a more proactive and innovative regulatory approach. Scholars in various sociotechnical fields advocate for transparency, inclusivity, and empathy in AI, discussing the importance of understanding and considering the human context in all phases of AI development (Ananny & Crawford, 2018; Madaio et al., 2020, von Eschenbach, 2021; Mosqueira-Rey et al., 2023) for Human-AI Interaction (HAII).

Having completed the first, foundational stage of the work with a full reconceptualization, the next will proceed explaining the state of the art in trustworthy AI frameworks and models, the result of a focused literature review

concluded in 2023. The following section concentrates on representative trustworthy AI proposals that fulfil some of the requirements of trustworthy AI and the technologies and sub technologies that back their design.

Trustworthy AI Frameworks and Models in the Literature - From Blockchain to Societal Integration

Several sociotechnical frameworks, strategies, and models have been developed in the past 5 years to enhance the trustworthiness of AI; each prioritising different aspects compatible with the conceptualisation synthesised in the previous section. However, many of these produce theoretical approaches, concentrate on a singular AI training method or a specific area of application. This sharply contrasts with three initiatives that by the time of completion of this work are noteworthy for their practical solutions to pressing research and policy challenges: the BF.TAI by Nassar et al. (2020, pp.7-10), the BlockIoTIntelligence by Singh et al. (2020, pp. 723-727), and the SITL by Rahwan (2018). While the BF.TAI framework addresses transparency issues, and the BlockIoTIntelligence architectural model increases data governance, and technical robustness and safety, the SITL consists of a formula to operationalise and augment human agency and oversight and diversity and non-discrimination.

Concretely, the BF.TAI and the BlockIoTIntelligence use BC technology for its advantages in data provenance (traceability), operational visibility, immutability (Sarpatwar et al., 2019, pp. 142-151), and decentralised nature (Ahmad et al., 2021, pp. 4-12). But despite addressing some robustness, lawfulness and ethicality, and human centredness concerns, these designs are not comprehensive and need a better alignment with the regulatory trustworthiness criteria.

The technical constituents of these models and how they operate will be delineated next, highlighting the elements that will be retained for the design of a fully compliant trustworthy AI model.

Blockchain Framework for Trustworthy AI

Nassar et al. (2020) proposed the BF.TAI to enhance the ´explainability´ of AI systems. The architecture of BF.TAI consists of Frontend Decentralized Applications (DApps), an Access Layer, an AI Layer, a Support Service and a BC Platform. The Frontend DApps are characterised by their decentralised and open-source interactive nature. They permit users inspections on the BC network (Wu, 2019, p. 1) and reliance on different forms of displays from command-line interfaces to mobile and web-based dashboards where the stakeholders can configure parameters and reach services.

The Access Layer, as the first component of the backend, facilitates Web3-based direct communication between the Frontend DApps and the BC platforms, utilizing the JavaScript Object Notation - Remote Procedure Call Application Programming Interface (JSON-RPC API) protocol for seamless data transfer from web-enabled applications to the Ethereum BC network. This layer extends its capabilities by incorporating conventional communication protocols such as REST HTTP for cloud data center connectivity, Java Message Service (JMS) APIs for intra-application communication, and Simple Object Access Protocols (SOAP) for sensor-based data sources. Supporting different data transfer protocols and allowing the JSON's RFC 4627 that has a lightweight data interchange format, the Access Layer ensures multi-level communication across various platforms and environments.

The AI Layer, at the heart of the framework, hosts AI and XAI predictors responsible for processing data and generating decision outcomes, these last being powerful forecasting algorithms.

The BF.TAI's Support Services handle registration and reputation management operations to oversee the ecosystem's stakeholders and maintain the integrity and reliability of the predictors. The BC Platform encompasses the BC network and decentralised storage solutions (Karaarslan & Konacaklı, 2020, pp. 57-58, 60, 63-64), adding the necessary infrastructure for running 'Smart Contracts'(SC) (Mik, 2017, pp. 272-277) and securely storing outcomes and metadata for training processes. Given the particularities of SCs, they are crucial for Nassar et al.'s framework's integrity and functionality. SCs are self-executing transactional mechanisms (Zou et al., 2021, p. 2086) that in this framework govern interactions, unify AI predictions, and manage the reputational system that makes explanations possible, while indicating the reliability and plausibility of these explanations. The BF.TAI is 'controlled' by Registration, Reputation, Aggregator and AI-task SCs. The first manages oracle and predictor registration; the second maintains predictor reputation scales; the third, aggregates decision outputs and administers consensus processes; and the fourth coordinates the execution of algorithmic tasks based on Service Level Agreement (SLA) parameters set by the users. SLAs document the terms and conditions agreed upon by the parties/agents, as confirmed and incorporated into the BF.TAI, and the parameters (e.g. latency, pricing, penalties) instruct the predictors.

The working flow of the BF.TAI begins with the users' engagement at the first layer, when the initial inputs such as parameters, AI/XAI predictor selection, and service requests are given. This step tailors the system's response to the users' specific needs and preferences. The data and commands then go through the Access Layer that guarantees their smooth transmission to the backend. This layer, specialised in handling multiple data transfer protocols, routes the data to the AI Layer where processing and decision-making are performed by the AI and XAI

predictors running simultaneously. Support Services operate in the background, overseeing critical aspects like users' registration and the reputational management components of the framework. These management processes are governed by the Reputation and Registration SCs in the support service. Next, the predictors interact with the BC platform by recording and logging their decisions on an immutable ledger, using decentralised storage systems. The Aggregation SC compares the decisions recorded and determines the correct outputs based on multiparty/multiagent decision-making protocols. These operations are what Nassar et al. call a 'consensus' process. At this point the predictors are screened to satisfy the requirements set by the users. Finally, the DApps receive the AI-task SC's results for iteration or approval, thus completing the cycle.

The technical components of the BF.TAI and its functional aspects are valuable because of the ingenious combination of capacities and its potential to set transparency criteria. Central to this framework are the SCs and their autonomy to govern the entire end-to-end blockchain-based AI system and subsystems, by managing interactions among the components. In addition, the decentralised storage of data is secure and provides traceability to stakeholders about data processing. Storing large amounts of data on the B would be expensive due to the various charges for processing transactions that may arise, compared to using decentralised storage solutions like the InterPlanetary File System (IPFS) (Alizadeh et al., 2020). Predictors, SCs, support services and DApps are the technical components of this framework that will be optimised and used in the design of blockchain-based AI model introduced in this paper.

In sum, Nassar et al. (2020) assume that the lack of explainability in AI decision-making, particularly in critical systems, is a fundamental cause of distrust towards AI technologies (Nassar et al., 2020, pp. 1-2). However, 'explainability' alone does not determine the broader concept of transparency, which itself is just one of seven key requirements for establishing trustworthiness in AI technologies. The other two qualities integral to transparency are 'traceability' and 'communication'. The BF.TAI could augment 'traceability' and 'communication' features but this is not explicitly emphasised. In addition, this framework does not concentrate on the human agency and oversight aspects. The proponents apply some human-centredness principles to the systems' development process but seem unconcerned about the role humans may play after the AI becomes operational. This is a limitation that impairs the evaluation of AI systems on whether their outputs will consistently adhere to the principles that the conceptual framework upholds.

BlockIoTIntelligence Architectural Model

Singh et al. (2020) combines two types of systems and subsystems; an architectural setup of Internet of Things (IoT) devices with four algorithmic levels and a BC platform for various applications. The authors aimed at securing automation in IoT systems and thus contributed with a proposal that meets the key requirements of trustworthy AI data governance, and technical robustness and safety, primarily. The architecture consists of The Device Intelligence, the Edge Intelligence, the Fog Intelligence, and the Cloud Intelligence. AI and BC applications are incorporated in all levels to verify and guarantee that IoT-based data are analysed and executed according to specific requirements. In addition, these applications record the processes in a traceable manner on a distributed ledger. Other components of the architecture are the IoT devices themselves, AI-enabled base stations, AI-enabled fog nodes and a blockchain-based cloud service. The IoT devices are interconnected physical objects in a network, ranging from household items to advanced industrial equipment, embedded with sensors and software to detect user presence and exchange data across various domains like homes, cities, vehicles, and hospitals (Pinheiro et al., 2019, p. 8). Base stations are network infrastructure components that enable wireless connection between the devices and the network by producing radio signals within specific coverage areas (Liu, 2022, pp. 24-26), and AI-enabled base stations are algorithms in the nodes of the architecture's B network, responsible for analysing traffic data from IoT devices and sensors and enhancing overall performance. Broadly speaking, nodes are intersection/connection points within a communication network that play an essential role in processing, storing, and transmitting data (Lewis, 2009, pp. 25-28 & 380-381; Elrom, 2019). The AI-enabled fog nodes identify the traffic flow in the IoT networks and are associated to the base stations at the edge. Finally, a blockchain-based cloud service is a storing and managing setup (computing services) for big data that in this case integrates the decentralised, secure, and transparent characteristics of BC technology (Murthy et al., 2020, pp. 205201- 205203).

BlockIoTIntelligence operates hierarchically but without a strict centralised control. It adopts a swarm operational approach at every level, which means that independent but loosely coupled agents perform complex tasks and self-organise to harness collective collaboration (Beni, 2020, p.792), with agents leading where and when needed.

The working flow starts at the Device Intelligence, involving numerous interconnected IoT devices and sensors, from simple data-gathering instruments to more complex appliances with their unique Frontend solutions (interfaces). The devices are tasked with the autonomous collection of raw data through AI learning processes that enhance the system continuously and securely due to the BC applications integrated at this level. The massive amount of row data collected ascend to the Edge Intelligence where the purpose is to analyse network issues via AI-enabled base stations, also connected to the BC. The base stations are responsible for the evaluation of the traffic data to assess critical network issues, including network scalability, and load balancing, thereby enhancing the AI system's technical robustness and safety. Each AI-enabled base station is in thus connected to sensing

devices.

Next is the reporting of the processes and data to the Fog Intelligence that runs algorithms for training models and optimal real time decision-making on the traffic flow of the raw data. At this point, the system considers resource management, energy consumption, scalability challenges, etc. via AI-enabled fog nodes. The BC technology within the Fog Intelligence establishes a distributed repository, ensuring that each node within the network keeps a copy of the entire ledger. This arrangement not only promotes data integrity and consistency but also enforces adherence to pre-defined rules and data transfer protocols within the IoT network. The Fog Intelligence layer, therefore, serves as a critical junction where powerful AI processing capacities, along with the security and decentralised benefits of BC technology, consolidate the capabilities and robustness of the IoT network.

At the top of the architecture is the Cloud Intelligence, where BC and AI for IoT applications converge in centres dedicated to protecting and curating the big data obtained from the IoT networks.

The integration of BC in every operational aspect of the BlockIoTIntelligence architectural model is valuable from a technical point of view, because it adds security, integrity and accuracy to the cloud intelligence performance, while boosting the big data analysis capacity for IoT applications. The proposal solves several well-known concerns in the field of IoT regarding data flows and governance. An added value of this design is the decentralisation of the network, because of the exponential speed in recording and sharing transactions and timestamps across nodes, obviating the need for third-party intermediaries. By maintaining data integrity throughout its lifecycle and countering the risks of centralisation, the BC technology increases the data governance capacity of the model. The hierarchical architecture also improves data governance through compartmentalisation, enabling specialised focus and scalability horizontally. The technical robustness and safety are heightened by the self-organising and meticulously coordinated swarm intelligence, influencing appropriate data handling and security against attacks. The design of the optimised blockchain-based AI model introduced in this paper will maintain this approach, as well as three of its components: the AI-enabled base stations, the AI-enabled fog nodes and the blockchain-based cloud service to extend automation to the data collection, analysis and storage.

Singh et al. designed the BlockIoTIntelligence, as a blockchain-enabled Internet of Things (IoT) architecture with an AI ‘model’ to decentralise big data analysis for IoT applications (Singh et al, 2020, p. 722). Ultimately, they tackled data governance, and technical robustness and safety challenges bringing together the BC platform and the processing capacities of AI. However, the critical role of human agency and oversight in achieving both was completely overlooked. Human agency and oversight are needed in the enforcement of structured protocols, for AI systems' resilience, and stability under varied conditions (cyber-risks), and conformity with the principles outlined in the regulatory framework.

Society-In-The-Loop (SITL) Framework

Rahwan (2018) draws on social contract theory to propose a sociotechnical framework that infuses societal values into AI's development and implementation. He speaks of an ‘algorithmic social contract’ resulting from the operationalisation of HITL through participation processes. Whereas ‘social contract’ refers to an implicit but foundational agreement between societal groups and their governing authorities (Rachels & Rachels, 2014, pp. 82-97), Rahwan’s algorithmic social contract represents terms and rules convened among human stakeholders and mediated by algorithms, that guarantee AI systems operate ethically, conforming to societal values and expectations. The components of the SITL are the AI system application, a consensus mechanism and a feedback loop. The specific AI application within the SITL would depend on the state and capabilities of AI technologies at any given point such as the systems and platforms available or in use. These may include existing machine learning models, data analysis tools, automated systems, and other forms of advanced computational arrangements that could be implemented or deployed in real-world scenarios, serving as hosts of the SITL framework. The framework fusions HITL and AI capabilities turning the consensus mechanism into a distinct component where humans are integral part. It involves society collectively deciding on the trade-offs between different values that AI may prioritise, such as balancing security with privacy or navigating various notions of fairness. Additionally, it encompasses the agreement on how the benefits and costs of AI systems are distributed among stakeholders. This mechanism mirrors the democratic processes in human governance, were public opinion influences voting and informs collective decision-making, adopting an approach more in line with human-centredness, the third characteristic of trustworthy AI restated earlier. The last component is a dynamic process denominated feedback loop where the identification of values and societal expectations (e.g., goals, ethics, and norms) takes place to later feed the system. This loop relies on adaptive learning capabilities that adjust and refine parameters based on evaluations against these human values and societal standards.

The framework may enhance virtually any AI system. It works in a cycle that begins forwarding outputs of the given AI application to the consensus mechanism, where ‘society’, represented by groups of stakeholders, must resolve contradictions, reconcile different values and weight gains and losses, such as balancing security and privacy or various notions of fairness. It also optimizes transactions costs and distributes the benefits and costs of automation. For illustration, consider the allocation of safety upgrades between passengers and pedestrians in driverless cars, and determining permissible collateral damage levels in conflicts using automated warfare systems. The

assessment outputs based on the collective input forming the algorithmic social contract is forwarded back to the initial AI system (the feedback loop) for re-training and alignment. This process requires adaptive learning, where the algorithm iteratively adjusts based on new data and evolving societal standards, ensuring its decisions remain ethically sound and representative of the collective understanding.

The SITL framework is valuable because it suggests a paradigm shift when prescribing the incorporation of collective ethics and societal values in AI's development, implementation and use. Rahwan emphasises the need for consensus and iterative feedback to give voice to the people in the assessment of the systems' outputs, in accord with the foundational rights of AI trustworthiness (democracy, justice and the rule of law, and citizens' rights), the human autonomy principle, and the human agency and oversight key requirement. While earlier proposals take HITL, HOTL, and HIC methods into account, Rahwan's stands out for its commitment to exploring societal influence on AI and its potential repercussions. It should be noted that this society-centric approach differs from initiatives involving the design of human/individual, computer/machine interaction strategies, or expert operational oversight mechanisms.

Some adjustments could enhance the usefulness of this framework. Firstly, running without a dealing mechanism on the assessment data depending on majority criteria, which reduces diversity, potentially leading to bias, discrimination, and underfitting during the re-training process (Martínez Ramil, 2021, p.5; Cunningham & Delany, 2021). Secondly, protecting individual interests at the HAII levels where the exercise of individual autonomy-freedom can truly be asserted. The SITL encourages collective ethics by considering a broad societal perspective and calling for AI systems to reflect social values and ethics. Even if conventions are not universally valid or static, this helps consolidate the current research agenda on the governance of AI systems that places societal and environmental wellbeing at the top. Moreover, engaging a wider spectrum of stakeholders may be seen as a precondition of inclusion and equitability. In this light, the author's socio-technical proposal points the essential challenge of aligning human and AI systems out, as highlighted by Hine & Floridi et al. (2022) later, but does not fully resolve it.

The consensus-driven decision pathway taken by the SITL proposal advances one trustworthy AI characteristics in a very limited manner, restricting the diversity of assessment data options/criteria. Rahwan's emphasis on stakeholder participation exclusively, downgrades the importance of 'expert', 'operational' -human- oversight, that are necessary, in conformity with the said key requirement, provided that the strengths of HITL, HOTL, and HIC are also incorporated. In doing so, the system would actively involve human decision-making in AI processes, maintaining direct control when needed. Blending these mechanisms to bring abstract concepts to practice in combination with the SITL idea would refine the governance of AI. This integration would cater to a multiplicity of applications and respect the complexity of human and societal values, creating sophisticated systems, adaptable to various contexts and evolution of societal norms.

In conclusion, the SITL framework articulates human agency and oversight but does not explore other regulatory components of trustworthiness as an attribute of AI.

The Blockchain-Based Trustworthy AI Model (BCTrustAI.SL) Design

BCTrustAI.SL is an original training and data analysis model that enhances the end-to-end trustworthiness of AI systems, comprehensively. It offers a trustworthy-by-design alternative which optimises earlier initiatives and recombines some of their features, surpassing their individual capacities. It consists of powerful technical aspects, components and principled processes that satisfy the regulatory criteria of trustworthiness discussed in section 3. The technical components pertain to the architectural tools (including technologies and sub-technologies) and methodologies that strengthen the reliability of automation systems.

Technical Components

Architecturally, the BCTrustAI.SL spreads over two tiers as Fig. 3 shows. The Backend has five intelligence layers for access, Smart Contracts (SCs), Human Agency and Oversight (HAO), the core AI subsystem, and data governance. This last is divided into two sub-layers with four levels each: an IoT Device Sub-layer receiving and managing data collected via IoT sources, and the Other Data Sub-layer that obtains the data from the rest of sources via web-based platforms (WBPs).

The engagement point for the stakeholders is at the Frontend Tier with its DApps components. These can be Command Line Interface-based (CLI-based), or interactive mobile or web-based dashboard applications that enable stakeholders to choose parameters with pre-coded SLAs for the HAO operations. The DApps should adhere to usability (UX) criteria, including the heuristics for graphic user-centric design (Jang et al., 2020, pp. 226215-226216 & 226220-226221) and legally relevant attribution factors, as classified by Solarte-Vasquez & Nyman-Metcalf (2017 pp. 226-232), these criteria play a dual role in affirming human-centredness principles and upholding transactional legitimacy.

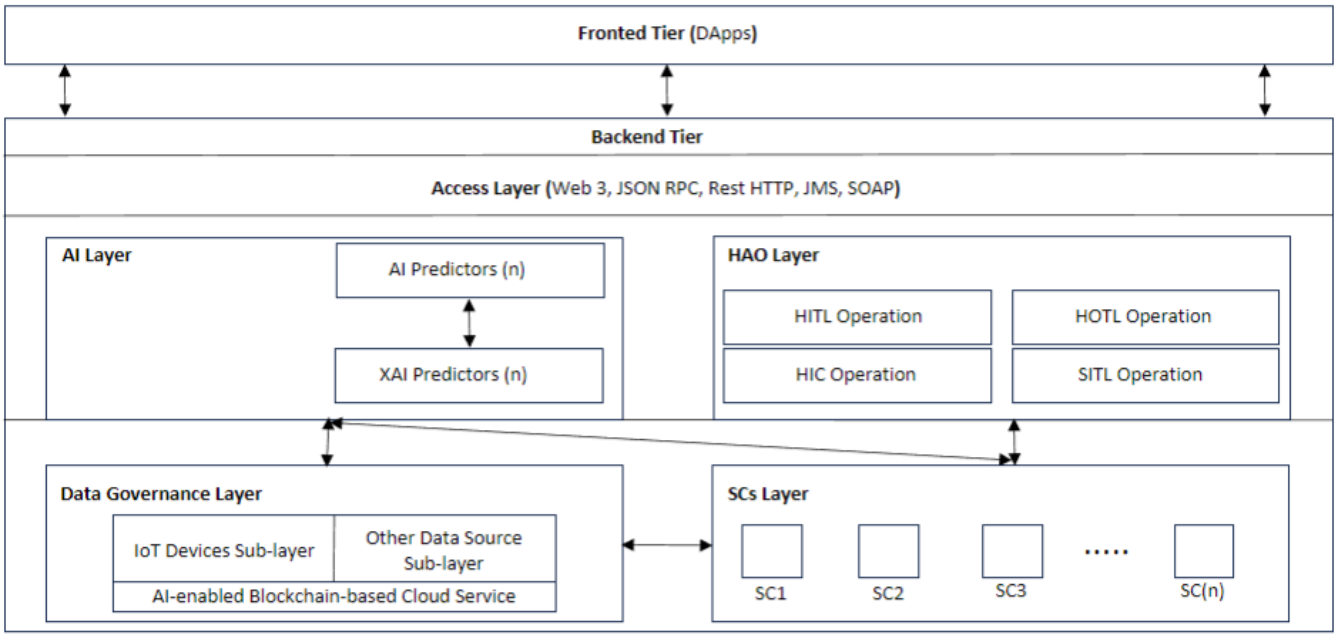

Fig. 3. Blueprint of BCTrustAI.SL.

In the Backend, the Access Layer functions as a bridge between the two tiers, creating reliable data links between the DApps and the SCs. It admits multiple data transfer protocols to increment the system's flexibility, and Web3 interfaces for direct communication between the DApps and the BC platform. Chief to its functioning is the use of JSON-RPC API, enabling data transfers between web-enabled applications and the Ethereum BC network. Transfers take place through remote procedure calls (RPC), utilising the lightweight and stateless JSON-RPC protocol that supports sockets, processes, HTTP, and other message-passing environments in JSON's RFC 4627 data format. The composition incorporates conventional communication protocols and APIs for connectivity, including Representational State Transfer Hypertext Transfer Protocols (REST HTTP) for cloud data center communications, JSON-RPC for client-server interactions with centralised repositories, Java Message Service (JMS) API for intra-application communication, and Simple Object Access Protocol (SOAP) for data exchange from the sensor-based data sources to the Backend Tier. Each of these components plays a vital role supporting the system’s interoperability and efficient data management in different environments and technologies.

The SCs layer is responsible for governing the system autonomously by conducting data transactions according to the parameters received from the SLAs, as in the BF.TAI framework. The amount of SCs and their duties are not definite. They are specifically tailored to the AI applications that are deployed or trained based on this model. This adaptability is integral to the design, allowing for the optimisation of both the quantity and functions of the SCs with each new implementation. This layer connects directly with the HAO, AI, and Data Governance Layers.

The HAO layer manages HITL, HOTL, HIC, and SITL processes. Consequently, impact assessments, user experience (UXI) evaluations and ethical compliance monitoring protocols are launched from here but having this layer does not exclude other HAO activities taking place throughout the model and deriving from a diversity of AI operations, for instance at the Frontend. This layer is inspired in the Support Services introduced by the BF.TAI framework.

In the AI Layer are the algorithms for data processing and decision-making, and XAI algorithms that inform how these processes take place. They substantiate the outcomes delivered by the system, like in Nassar et al.’s framework. The number of predictors is adjustable depending on the AI application deployed or trained according to the BCTrustAI.SL.

The Data Governance Layer controls the storage and management of the system data (input, output and insights from the training and analysis, and the operational and raw data gathered during collection phases). The layers are structured hierarchically, as Fig. 4 shows, with levels and sub-layers which have the primary function of gathering data; one is specialised in IoT sources, while the other in WBPs such as social media platforms, public and governmental databases, surveys, polls, corporate and scientific research repositories, etc. The two are built to handle these inputs securely. The levels are the IoT devices and Web, Edge, Fog, and Blockchain-based Cloud Service. This last is shared by both sub-layers, as a singular component.

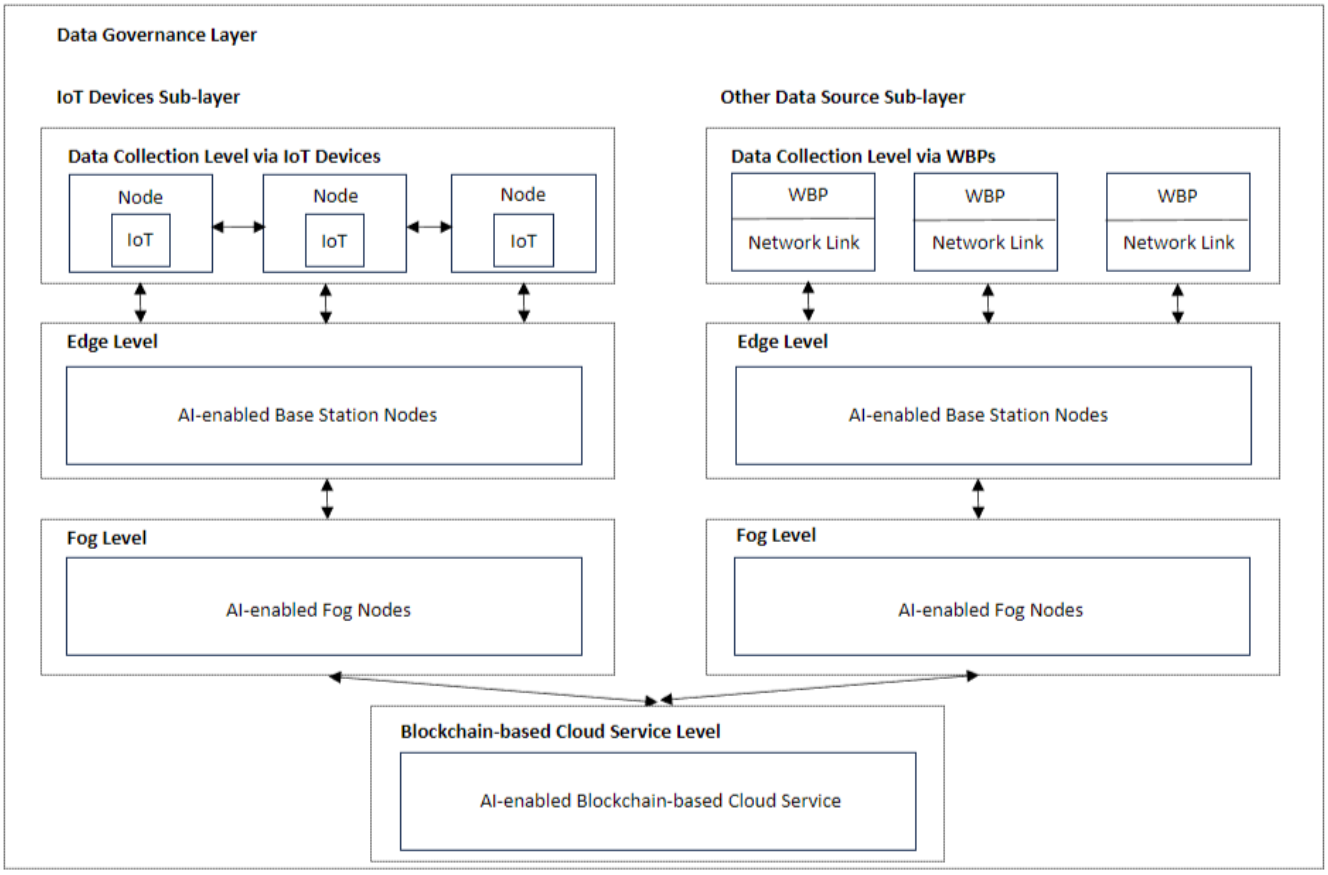

Fig. 4. Detailed Data Governance Layer of the BCTrustAI.SL

The data entry point is at the collection level of the sub-layers. On the one hand, the Data Collection Level via IoT Devices, relies on the IoT devices themselves and nodes like Dammak et al. (2022) suggests in their architectural model. The nodes create a sub-network in the BC network and expedite communication between devices. On the other hand, the Other Data Collection Level extracts data from WBPs and through the Network Links of the model, which requires a wide selection of data transfer protocols corresponding to the Access Layer setup. The protocols work like linkages that make web platforms and the Web 3 environment of the BCTrustAI.SL compatible. The Edge Level hosts the Intelligent base station nodes, which gather and classify external traffic data in the way the BlockIoTIntelligence architectural model does. The Fog Level groups the Fog Nodes that are critical components in the execution of algorithms handling resource management, energy efficiency, and scalability, similarly to the Fog nodes of the BlockIoTIntelligence framework as well. This is performed on the external traffic data in real-time. The last is the common Blockchain-based Cloud Service Level, backed by its algorithmic intelligence. AI augments the efficiency of the system, while BC technology adds robustness and security to the distributed storage system.

Workflow Dynamics

In practice, the BCTrustAI.SL system activates when stakeholders, primarily experts (developers and/or handlers/users depending on the system) engage with the DApps at the Frontend Layer. These interactions are a refinement of the BF.TAI in that they are more open. They choose the parameters in the pre-set SLAs, aligning to their goals, which trigger the relevant SCs that execute and manage all algorithmic processes. The SCs initiate the data collection, training and analysis processes, some of which occur at the two sub-layers of the Data Governance Layer, while others at the AI Layer, using AI and XAI predictors. The SLAs are predefined in the sense they work as customisable templates to meet the unique terms and conditions of each operation that the SCs must run. The templates are multimodal and should be prepared by the developers, based on expert knowledge, representing different types of operations such as criteria to produce a medical diagnosis, social media data analysis for scoring, initial data processing, outputs' assessment according to UXI factors, legal or ethical compliance, etc. The phases and workflow of the model can be seen in Fig. 5.

The Data Governance Layer in BCTrustAI.SL broadens the capacity and resourcefulness of the BlockIoTIntelligence architectural model to obtain data. The first raw data input at the layer comes from IoT Devices and WBPs. In the first sub-layer, a diverse array of interconnected IoT devices and sensors linked in nodes produce raw data that are sent to the corresponding Edge Level. The nodes work in swarm intelligence mode. AI-enabled Base Station Nodes use the network formed by IoT devices to collect the second input, which consists of external traffic data. These, together with the kind sourced from WBPs are subject of Edge Level algorithmic evaluations regarding scalability, load balancing and other critical network issues. The goal of this performance assessment is

to check the efficiency of the network. In the other sub-layer, the raw and external traffic data are also delivered from the WBPs through the Network Link of the Data Collection to the respective Edge Level. The two sub layers may work on their own or simultaneously, depending on the application and the ML training expected (determined by the parametric specifications established in the first step). The first raw data input of the two sub layers moves to their Fog Levels where AI-enabled fog nodes analyse the internal traffic of the that input in real time, and respond to challenges about resource management, energy efficiency, scalability, latency, interoperability, data integrity, cost management, fault tolerance etc. Next, all this input and the first set of output data from the traffic analysis at each sub-layer are transmitted to the Blockchain-based Cloud Service Level where they are sorted, pre-processed and stored. At this point all data has gone though some degree of processing and can no longer be considered 'raw'. The results of this step excluded the traffic related data will serve as the main input for the AI Layer where the core AI operations take place, be them for data analysis or ML training purposes. The data flows between the AI-enabled Blockchain-Based Cloud Service and the AI and XAI predictors in a synchronous and continuous exchange. These explainability techniques and the link with the Data Governance Layer resemble the components and operations of the BF.TAI, but BCTrustAI.SL has advantages. The main input data is forwarded to the AI Layer for algorithm training and analysis; a preliminary output is transmitted back to the Cloud Service Level at the Data Governance Layer along with the insights created by the predictors for a. storage and b. extra HAO activities involving expert stakeholders and users. Finally, the combined preliminary output and insight data are sent to the HAO Layer for appraisal and/or other HAO operations such as ethical evaluations, UX/UXI screening, legal compliance monitoring, etc. The HAO Layer and post-training flows are inspired in the SITL framework, discarding the inexplicit consensus-driven decision-making pathway. Instead, the CCTrustAI.SL design introduces complementing HITL, HOTL, HIC and SITL properties.

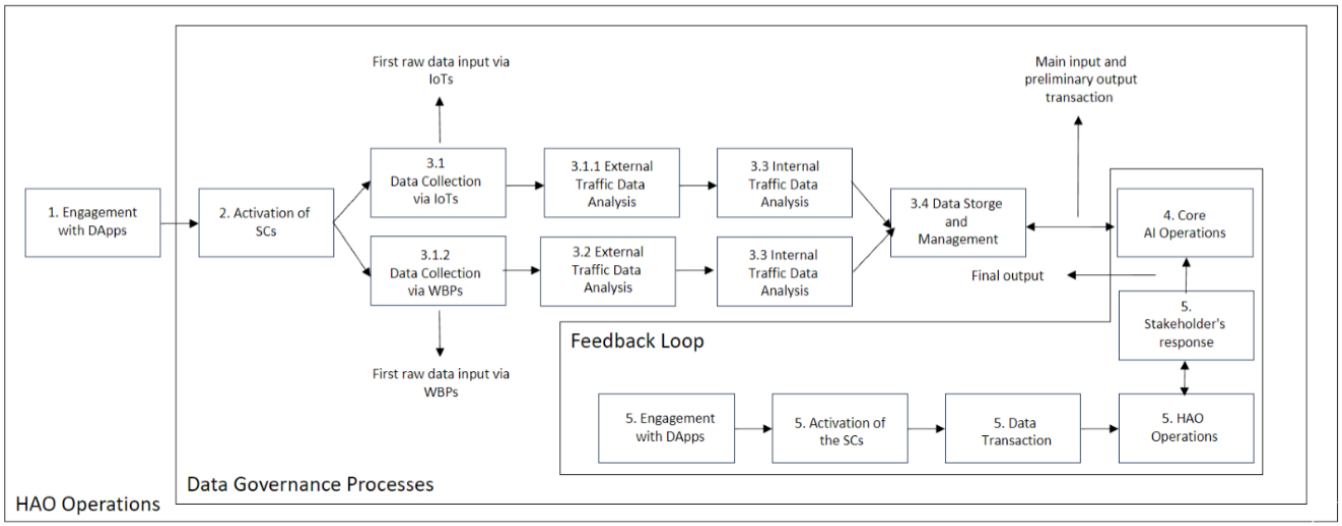

Fig. 5. Operational Workflow of BCTrustAI.SL.

These last steps require the stakeholders to interact with DApps and define SLAs, based on a different set of templates for post-analysis/training HAO operations. The purpose is to activate the SCs that manage transactions between the layers. Ultimately, decision-making responsibilities belong to the stakeholders who choose whether to accept, decline or request improvements. In fact, these cycles describe a feedback loop subsystem that implements and extends the initiative of the SITL framework. The training phases will restart when final outputs are declined, and the HAO loops will repeat until adjustments are no longer necessary.

Resulting Conformity of the BCTrustAI.SL With the Emerging Institutional Background

BCTrustAI.SL attains the operationalisation of abstract regulatory conditions, by building a complex but solid infrastructure that integrates selected technologies and principled processes. The model facilitates a comprehensive alignment with the three widely established characteristics defining the trustworthiness of AI: robustness, ethicality and lawfulness, and human-centredness. More specifically, the novelty of BCTrustAI.SL is advancing the system's data governance and HAO beyond the state of the art, meticulously integrating human centric processes into every step of the model's operational workflow.

The BCTrustAI.SL architecture is robust because it ensures durability and operational efficiency and protection against cyber risk challenges by meeting the respective key requirements. Data governance is a precondition of the rest because it allows incident response planning and control over the data quality, audits and monitoring processes adhering to regulations like the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679), risk management, etc. In the model, data governance transcends the mere handling of raw data, extending to all kinds of inputs and outputs produced during the operational timeline/phases of the system/AI

application. This optimisation of the BlockIoTIntelligence architectural model will develop trust in the systems and ease the uptake and use of AI applications. Hence, the Data Governance Layer is much more than a data repository, it is intended to be a comprehensive management strategy for the data and the working flows within the system. This approach exploits the most notable BC technology qualities, such as traceability, to build on several forms of HAO for control and monitoring, making them available at all times. HAO opportunities and allowances are essential for effective governance activities like compliance oversight in the collection, management and usage of data. Basically, various stakeholders play the role of protecting the data's integrity from unauthorised access or manipulation. HAO processes also generate cyber security and data protection protocols to strengthen resilience against adversarial attacks and data breaches, thereby reinforcing the overall robustness of the AI system in accordance with the harm prevention principle. Even though BC technology's inherent immutability impedes manipulation attempts, the omnipresent control mechanisms of BCTrustAI.SL leave no doubt about the model's observance of the technical robustness and safety, and data protection key requirements. In addition, the swarm intelligence approach is geared to augment the efficiency and technical robustness and safety when the input sources are IoT devices.

Regarding ethicality and lawfulness, the design of the model satisfies its three key requirements: transparency, accountability, and diversity and non-discrimination. First, the BCTrustI.SL accomplishes a significant level of transparency combining BC technology and algorithmic intelligence like in Nassar et al.'s framework (2020) but highlighting the role of 'traceability' and 'communication' as much as they do 'explainability'; i.e., the first engagement process refines the BF.TAI to gain on the 'communication' sub-requirement. Because transparency is a precondition of the other two features, the optimised model aims at enhancing the accurate interpretation of outputs with HAO operations, such as the feedback loop and risk identification steps/activities, for example, checks on inclusion and neutrality of the systems' preliminary outputs. At the same time, transparency supports human agency and oversight increasing accessibility and greater awareness about the system's functioning, which may foster engagement and empowerment (HIC). As explained above, the data governance strategies introduced by the BF.TAI framework are acknowledged as especially helpful contribution to the operationalisation of trustworthiness and thus, the design implements a comparable approach benefitting from the aggregation of internal data (explanations issued by XAI predictors in the AI Layer) and external data, derived from the stakeholders' interpretation (main input, preliminary output, and parameters in the HAO operations). This dual approach addresses most 'explainability' questions, especially considering the procedural improvements that make data governance possible. Last, the capacity to carry HAO operations facilitating human engagement throughout the system conforms by far with the 'communication' sub-requirement, enhancing the openness of the system for the scrutiny and accountability of AI/ML applications.

BCTrustAI.SL raises the human-centredness bar by not only meeting basic regulatory conditions but also expanding the understanding and reach of the concept and demonstrating its horizontal importance to any trustworthy AI system. To satisfy the human agency and oversight key requirement, the model habilitates as many human interaction opportunities as necessary, from start to end of the workflow, especially during training and retraining cycles. BCTrustAI.SL, administers HITL, HOTL, HIC, and SITL transactions in one of its four layers, prompting human engagement in the Fronted Tiers and active participation in the evaluation of outcomes. The interfaces may also be augmented by the model, which means that the applications where auditing entails UXI evaluation will produce improvements. UXI testing is fundamental because all HAO operations must be informed to reduce errors and guarantee procedural and substantive -legal- validity. A good interface design reduces frictions in communication when HAII interactions are inviting, responsive and conversational for the users/operators. It stages the first point of contact with the AI application and affects human participation, motivation, comprehension and ultimately, trust. In general, to facilitate monitoring of AI/ML training and retraining, the human factor should not be dismissed, but most particularly, if systems shall remain tools that require constructive human judgement.

In summary, the proposed model operationalises the trustworthy AI concept fulfilling the extant regulatory trustworthiness criteria for AI applications and the emerging institutional background. Its uniqueness stems from its capability to operationalise robustness, ethicality and lawfulness and human-centredness, and from its readiness to proceed with the proof of the concepts neighbouring human centredness i.e. HITL, HOTL, HIC, and SITL in automation as HCI and HAII design problems.

Resulting Conformity of the BCTrustAI.SL With the Emerging Institutional Background

BCTrustAI.SL attains the operationalisation of abstract regulatory conditions, by building a complex but solid infrastructure that integrates selected technologies and principled processes. The model facilitates a comprehensive alignment with the three widely established characteristics defining the trustworthiness of AI: robustness, ethicality and lawfulness, and human-centredness. More specifically, the novelty of BCTrustAI.SL is advancing the system's data governance and HAO beyond the state of the art, meticulously integrating human centric processes into every step of the model's operational workflow.

The BCTrustAI.SL architecture is robust because it ensures durability and operational efficiency and protection

against cyber risk challenges by meeting the respective key requirements. Data governance is a precondition of the rest because it allows incident response planning and control over the data quality, audits and monitoring processes adhering to regulations like the General Data Protection Regulation (GDPR) (EU2016/679), risk management, etc. In the model, data governance transcends the mere handling of raw data, extending to all kinds of inputs and outputs produced during the operational timeline/phases of the system/AI application. This optimisation of the BlockIoTIntelligence architectural model will develop trust in the systems and ease the uptake and use of AI applications. Hence, the Data Governance Layer is much more than a data repository, it is intended to be a comprehensive management strategy for the data and the working flows within the system. This approach exploits the most notable BC technology qualities, such as traceability, to build on several forms of HAO for control and monitoring, making them available at all times. HAO opportunities and allowances are essential for effective governance activities like compliance oversight in the collection, management and usage of data. Basically, various stakeholders play the role of protecting the data's integrity from unauthorised access or manipulation. HAO processes also generate cyber security and data protection protocols to strengthen resilience against adversarial attacks and data breaches, thereby reinforcing the overall robustness of the AI system in accordance with the harm prevention principle. Even though BC technology's inherent immutability impedes manipulation attempts, the omnipresent control mechanisms of BCTrustAI.SL leave no doubt about the model's observance of the technical robustness and safety, and data protection key requirements. In addition, the swarm intelligence approach is geared to augment the efficiency and technical robustness and safety when the input sources are IoT devices.

Regarding ethicality and lawfulness, the design of the model satisfies its three key requirements: transparency, accountability, and diversity and non-discrimination. First, the BCTrustI.SL accomplishes a significant level of transparency combining BC technology and algorithmic intelligence like in Nassar et al.'s framework (2020) but highlighting the role of 'traceability' and 'communication' as much as they do 'explainability'; i.e., the first engagement process refines the BF.TAI to gain on the 'communication' sub-requirement. Because transparency is a precondition of the other two features, the optimised model aims at enhancing the accurate interpretation of outputs with HAO operations, such as the feedback loop and risk identification steps/activities, for example, checks on inclusion and neutrality of the systems' preliminary outputs. At the same time, transparency supports human agency and oversight increasing accessibility and greater awareness about the system's functioning, which may foster engagement and empowerment (HIC). As explained above, the data governance strategies introduced by the BF.TAI framework are acknowledged as especially helpful contribution to the operationalisation of trustworthiness and thus, the design implements a comparable approach benefitting from the aggregation of internal data (explanations issued by XAI predictors in the AI Layer) and external data, derived from the stakeholders' interpretation (main input, preliminary output, and parameters in the HAO operations). This dual approach addresses most 'explainability' questions, especially considering the procedural improvements that make data governance possible. Last, the capacity to carry HAO operations facilitating human engagement throughout the system conforms by far with the 'communication' sub-requirement, enhancing the openness of the system for the scrutiny and accountability of AI/ML applications.

BCTrustAI.SL raises the human-centredness bar by not only meeting basic regulatory conditions but also expanding the understanding and reach of the concept and demonstrating its horizontal importance to any trustworthy AI system. To satisfy the human agency and oversight key requirement, the model habilitates as many human interaction opportunities as necessary, from start to end of the workflow, especially during training and retraining cycles. BCTrustAI.SL, administers HITL, HOTL, HIC, and SITL transactions in one of its four layers, prompting human engagement in the Fronted Tiers and active participation in the evaluation of outcomes. The interfaces may also be augmented by the model, which means that the applications where auditing entails UXI evaluation will produce improvements. UXI testing is fundamental because all HAO operations must be informed to reduce errors and guarantee procedural and substantive -legal- validity. A good interface design reduces frictions in communication when HAII interactions are inviting, responsive and conversational for the users/operators. It stages the first point of contact with the AI application and affects human participation, motivation, comprehension and ultimately, trust. In general, to facilitate monitoring of AI/ML training and retraining, the human factor should not be dismissed, but most particularly, if systems shall remain tools that require constructive human judgement.

In summary, the proposed model operationalises the trustworthy AI concept fulfilling the extant regulatory trustworthiness criteria for AI applications and the emerging institutional background. Its uniqueness stems from its capability to operationalise robustness, ethicality and lawfulness and human-centredness, and from its readiness to proceed with the proof of the concepts neighbouring human centredness i.e. HITL, HOTL, HIC, and SITL in automation as HCI and HAII design problems.

Conclusion

This paper reported on a two-stage, multimodal study, each contributing to the operationalisation of the normative discourse on trustworthy AI. The first stage synthesised the emerging regulatory notion of trustworthiness to make a proper delimitation in favour of its institutionalisation as a chief attribute of AI. It did so, following a partial institutional analysis that involved formal sources and the result of a literature review (public policies, regulatory

documents, and scholarly literature focusing on research published by organisations, practitioners, and experts), unpacking the features of trustworthy AI.

The second stage built on these results, added a revision of the state of the art, which revealed gaps in the technical literature. It presented the BCTrustAI.SL, a trustworthy-by-design sociotechnical model that translates the abstract concepts from the policy discourse into components, principled processes and operations. A simplified AI/ML model design method allowed, at a conceptual level, the introduction of this proposal that facilitates human-machine collaboration while aligning with current regulatory standards.

The BCTrustAI.SL retains components and operations from earlier proposals (engagement with DApps, activation SCs via SLAs, data governance, core AI operations via AI and XAI predictors and feedback loop) but surpasses their capacities in terms of legal and ethical conformity. Technically, the model recognises the advantages of BC (data provenance, traceability, operational visibility, data immutability among others) and puts this technology and its sub-technologies to use, in combination with AI (SC nodes, DApps, blockchain-based cloud service, decentralised computing, etc.). Operationally, the gains are accomplished with processes inspired in principles:

- It upgrades human agency and oversight and effective 'communication' for transparency by engaging the stakeholders in data governance processes, enriching the 'traceability' of the system.
- It expands the stakeholder's involvement from the limited HIC responsibilities monitoring the technical robustness of AI, thereby facilitating HAII and seamless collaboration between humans and the machine.
- It opens up to a broad but focused spectrum of stakeholders in post-training evaluations without relying on 'running code' (Calliess et al., 2010, pp. 134-153) to determine the validity of a social contract on high stake AI applications. Unlike the consensus-driven decision-making seen in SITL, BCTrustAI.SL permits a more inclusive approach.

Furthermore, the model augments the overall interaction between the stakeholders and the system. This is pressing in terms of procedural fairness and therefore invites towards the future, legally relevant UX/UXI adjustments which are deemed compelling in responsive and responsible automation.

BCTrustAI.SL, may mark a significant advancement in the realm of AI trustworthiness, but it could face implementation challenges. It still needs pre-set SLA templates tailored specifically to various HAO operations and its effective/lawful implementation cannot proceed without understanding the competences and needs of the stakeholders. The readiness of the model for a proof of concept depends on further performance/compliance/conformity-oriented research to create the procedural options that help AI meet the trustworthiness key requirements and uphold the model's dedication to human-centred and reliable AI development.

References

Ahmad, R. W., Hasan, H., Jayaraman, R., Salah, K., & Omar, M. (2021). Blockchain applications and architectures for port operations and logistics management. *Research in Transportation Business & Management*, 41, 100620. https://doi.org/10.1016/j.rtbm.2021.100620

Alizadeh, M., Andersson, K., & Schelén, O. (2020). Efficient Decentralized Data Storage Based on Public Blockchain and IPFS. 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 1–8. https://doi.org/10.1109/CSDE50874.2020.9411599

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. https://doi.org/10.1177/1461444816676645

Antonov, A., & Kerikmäe, T. (2020). Trustworthy AI as a Future Driver for Competitiveness and Social Change in the EU. In D. Ramiro Troitiño, T. Kerikmäe, R. M. de la Guardia, & G. Á. Pérez Sánchez (Eds.), The EU in the 21st Century: Challenges and Opportunities for the European Integration Process (pp. 135–154). Springer International Publishing. https://doi.org/10.1007/978-3-030-38399-2_9

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). *Synthesizing Robust Adversarial Examples.* arXiv. https://doi.org/10.48550/arXiv.1707.07397

Atoum, I., & Keshta, I. (2021). Big data management: Security and privacy concerns. *International Journal of Advanced and Applied Science*, 8, 73–83. https://doi.org/10.21833/ijaas.2021.05.009

Beni, G. (2020). Swarm Intelligence. In M. Sotomayor, D. Pérez-Castrillo, & F. Castiglione (Eds.), Complex Social and Behavioral Systems: Game Theory and Agent-Based Models (pp. 791–818). Springer US. https://doi.org/10.1007/978-1-0716-0368-0_53

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77–91.

Cachat-Rosset, G., & Klarsfeld, A. (2023). Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines. *Applied Artificial Intelligence*, 37(1), 2176618. https://doi.org/10.1080/08839514.2023.2176618

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Calliess, G.-P., Zumbansen, P., & Scott, C. M. (2010*). Rough Consensus and Running Code: A Theory of Transnational Private Law*. Hart Publishing.

China Academy of Information and Communications Technology (CAICT) & JD Explore Academy. (2021). White Paper on Trustworthy Artificial Intelligence.

Cunningham, P., & Delany, S. J. (2021). Underestimation Bias and Underfitting in Machine Learning. In F. Heintz, M. Milano, & B. O'Sullivan (Eds.), Trustworthy AI - Integrating Learning, Optimization and Reasoning (pp. 20–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-73959-1_2

Dammak, B., Turki, M., Cheikhrouhou, S., Baklouti, M., Mars, R., & Dhahbi, A. (2022). LoRaChainCare: An IoT Architecture Integrating Blockchain and LoRa Network for Personal Health Care Data Monitoring. *Sensors*, 22(4), Article 4. https://doi.org/10.3390/s22041497

Das, V., herukuri, A. K., Hu, Q., Kamalov, F., & Jonnalagadda, A. (2023). Proactive AI Enhanced Consensus Algorithm with Fraud Detection in Blockchain. In Y. Maleh, M. Alazab, & I. Romdhani (Eds.), Blockchain for Cybersecurity in Cyber-Physical Systems (pp. 259–274). Springer International Publishing. https://doi.org/10.1007/978-3-031-25506-9_13

Department of International Cooperation Ministry of Science and Technology (MOST), P.R. China. (2017). Next Generation Artificial Intelligence Development Plan Issued by State Council | China's Strengths Creates Innovation Miracles.

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer Int Publishing. https://doi.org/10.1007/978-3-030-30371-6

Dignum V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, 20(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z

Elrom, E. (2019). Blockchain Nodes. *The Blockchain Developer: A Practical Guide for Designing, Implementing, Publishing, Testing, and Securing Distributed Blockchain-based Projects* (pp. 31–72). Apress. https://doi.org/10.1007/978-1-4842-4847-8_2

European Commission. (2020). White Paper on Artificial Intelligence: A European approach to excellence and trust.

European Commission. (2021). Coordinated Plan on Artificial Intelligence 2021 Review.

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council. Official Journal of the European Union, L119, 1-88.

Executive Office of the President, Office of Management and Budget. (2020). Guidance for Regulation of Artificial Intelligence Applications (M-21-06).

Fanni R., Steinkogler, V. E., Zampedri, G., & Pierson, J. (2023). Enhancing human agency through redress in Artificial Intelligence Systems. *AI & SOCIETY*, 38(2), 537–547. https://doi.org/10.1007/s00146-022-01454-7

Felzmann H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, 26(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Floridi, L., Cowls, J., ltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples* (arXiv:1412.6572). arXiv. https://doi.org/10.48550/arXiv.1412.6572

Göksal, Ş. İ., Solarte Vasquez, M. C., & Chochia, A. (in press). The EU AI Act's alignment within European Union's regulatory framework on artificial intelligence. *International and Comparative Law Review*. Advance online publication.

Ishida, T. (2004). Society-Centered Design for Socially Embedded Multiagent Systems. In M. Klusch, S. Ossowski, V. Kashyap, & R. Unland (Eds.), Cooperative Information Agents VIII (pp. 16–29). Springer. https://doi.org/10.1007/978-3-540-30104-2_2

Independent High-Level Expert Group on Artificial Intelligence Set Up by the European Commission. (2019). Ethics guidelines for trustworthy AI. Brussels.

Jang, H., Han, S. H., & Kim, J. H. (2020). User Perspectives on Blockchain Technology: User-Centered Evaluation and Design Strategies for DApps. *IEEE Access*, 8, 226213–226223. https://doi.org/10.1109/ACCESS.2020.3042822

Janssen, M., rous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. *Government Information Quarterly*, 37(3), 101493. https://doi.org/10.1016/j.giq.2020.10149

Joamets, K., & Vasquez, M. C. S. (2020). Regulatory Framework of the Research-Based Approach to Education in the EU. *TalTech Journal of European Studies*, 10(3), 109–136. https://doi.org/10.1515/bjes-2020-0024

Jobin, A., Ienca, M., & ayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hine E., & Floridi, L. (2022). Artificial intelligence with American values and Chinese characteristics: A comparative analysis of American and Chinese governmental AI policies. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01499-8

Hohma, E., & Lütge, C. (2023). From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. *AI*, 4(4), Article 4. https://doi.org/10.3390/ai4040046

Karaarslan, E., & Konacaklı, E. (2020). *Data Storage in the Decentralized World: Blockchain and Derivatives* (pp. 37–69). https://doi.org/10.26650/B/ET06.2020.011.03

Kayser-Bril, N. (2020). Google apologizes after its Vision AI produced racist results. AlgorithmWatch. https://algorithmwatch.org/en/google-vision-racism/

Krishnan, M. (2020). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology*, 33(3), 487–502. https://doi.org/10.1007/s13347-019-00372-9

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & usbit, D. (2019). Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 182:1-182:26. https://doi.org/10.1145/3359284

Lehner, O. M., Ittonen, K., Silvola, H., Ström, E., & Wührleitner, A. (2022). Artificial intelligence based decision-making in accounting and auditing: Ethical challenges and normative thinking. *Accounting, Auditing & Accountability Journal*,

35(9), 109–135. https://doi.org/10.1108/AAAJ-09-2020-4934

Lewis, T. G. (2009). *Network Science: Theory and Applications* (1st edition). Wiley.

Li, N., Adepu, S., Kang, E., & Garlan, D. (2020). Explanations for human-on-the-loop: A probabilistic model checking approach. Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, 181–187. https://doi.org/10.1145/3387939.3391592

Liu, X. (2022). Chapter 1—Introduction. In X. Liu (Ed.), Optical Communications in the 5G Era (pp. 1–28). *Academic Press*. https://doi.org/10.1016/B978-0-12-821627-9.00012-7

M. Bublitz, F., Oetomo, A., S. Sahu, K., Kuang, A., X. Fadrique, L., E. Velmovitsky, P., M. Nobrega, R., & P. Morita, P. (2019). Disruptive Technologies for Environment and Health Research: An Overview of Artificial Intelligence, Blockchain, and Internet of Things. *International Journal of Environmental Research and Public Health*, 16(20), Article 20. https://doi.org/10.3390/ijerph16203847

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3313831.3376445

Marcus, G. (2018). *Deep Learning: A Critical Appraisal* (arXiv:1801.00631). arXiv. https://doi.org/10.48550/arXiv.1801.00631

Ramil, P. (2021). Is the EU human rights legal framework able to cope with discriminatory AI? *IDP: Revista de Internet, Derecho y Política = Revista d'Internet, Dret i Política*, Extra 34, 5.

Meurisch, C., Mihale-Wilson, C. A., Hawlitschek, A., Giger, F., Müller, F., Hinz, O., & Mühlhäuser, M. (2020). Exploring User Expectations of Proactive AI Systems. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(4), 146:1-146:22. https://doi.org/10.1145/3432193

Meurisch, C., & Mühlhäuser, M. (2021). Data Protection in AI Services: A Survey. *ACM Computing Surveys*, 54(2), 40:1-40:38. https://doi.org/10.1145/3440754

Mik, E. (2017). Smart contracts: Terminology, technical limitations and real-world complexity. *Law, Innovation and Technology*, 9(2), 269–300. https://doi.org/10.1080/17579961.2017.13784

Murthy, Ch. V. N. U. B., Shri, M. L., Kadry, S., & Lim, S. (2020). Blockchain Based Cloud Computing: Architecture and Research Challenges. *IEEE Access*, 8, 205190–205205. https://doi.org/10.1109/ACCESS.2020.3036812

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., obes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Nassar, M., Salah, K., ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, 10(1), e1340. https://doi.org/10.1002/widm.1340

National Institute of Standards and Technology, US Department of Commerce. (2022). AI Risk Management Framework: Second Draft.

Novelli C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01635-y

O'Neil (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (4th edition). Penguin Books.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information* (Reprint edition). Harvard University Press.

Pinheiro, A. J., de M. Bezerra, J., Burgardt, C. A. P., & Campelo, D. R. (2019). Identifying IoT devices and events based on packet length from encrypted traffic. *Computer Communications*, 144, 8–17. https://doi.org/10.1016/j.comcom.2019.05.012

Rachels, J., & Rachels, S. (2014). *The Elements of Moral Philosophy* (8th edition). McGraw Hill.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., edayatnia, B., Cheng, M., Nagar, A., King, E., Bland, K., Wartick, A., Pan, Y., Song, H., Jayadevan, S., Hwang, G., & Pettigrue, A. (2018). *Conversational AI: The Science Behind the Alexa Prize* (arXiv:1801.03604). arXiv. https://doi.org/10.48550/arXiv.1801.03604

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance), 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance) (2024). http://data.europa.eu/eli/reg/2024/1689/oj/eng

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd edition). Pearson.

Russell, S., Dewey, D., & Tegmark, M. (2016). *Research Priorities for Robust and Beneficial Artificial Intelligence* (arXiv:1602.03506). arXiv. https://doi.org/10.48550/arXiv.1602.03506

Sarpatwar, K., Vaculin, R., Min, H., Su, G., Heath, T., Ganapavarapu, G., & Dillenberger, D. (2019). Towards Enabling Trusted Artificial Intelligence via Blockchain. In S. Calo, E. Bertino, &D. Verma (Eds.), Policy-Based Autonomic Data Governance (pp. 137–153). Springer International Publishing. https://doi.org/10.1007/978-3-030-17277-0_8

Select Committee on Artificial Intelligence of the National Science and Technology Council. (2023). National artificial intelligence research and development strategic plan: 2023 Update.

Simonite, T. (2018). When It Comes to Gorillas, Google Photos Remains Blind. Wired. [online] Available at: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/ [Accessed 1 November 2024].

Singh S. K., Rathore, S., & Park, J. H. (2020). BlockIoTIntelligence: A Blockchain-enabled Intelligent IoT Architecture with

Artificial Intelligence. *Future Generation Computer Systems*, 110, 721–743.https://doi.org/10.1016/j.future.2019.09.002

Smuha, N. A. (2021). From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. *Law, Innovation and Technology*, 13(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Solarte-Vásquez, M. C., & Nyman-Metcalf, K. (2017). Smart Contracting: A Multidisciplinary and Proactive Approach for the EU Digital Single Market. *TalTech Journal of European Studies*, 7(2), 208–246. https://doi.org/10.1515/bjes-2017-0017

Sutton, R. S., Barto, A. G., & Bach, F. (2018). *Reinforcement Learning: An Introduction (second edition)*. MIT Press.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks* (arXiv:1312.6199). arXiv. https://doi.org/10.48550/arXiv.1312.6199

Taghikhah, F., Erfani, E., Bakhshayeshi, I., Tayari, S., Karatopouzis, A., & Hanna, B. (2022). Chapter 5 - Artificial intelligence and sustainability: Solutions to social and environmental challenges. In M. Asadnia, A. Razmjou, & A. Beheshti (Eds.), Artificial Intelligence and Data Science in Environmental Sensing (pp. 93–108). Academic Press. https://doi.org/10.1016/B978-0-323-90508-4.00006-X

Taylor, R. R., O'Dell, B., & Murphy, J. W. (2023). Human-centric AI: Philosophical and community-centric considerations. *AI & Society*. https://doi.org/10.1007/s00146-023-01694-1

The White House. (2023). FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.

United States Department of Defense. (2022). U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway.

Umbrello, S. (2021). Ai Winter. In M. Klein & P. Frana (Eds.), Encyclopedia of Artificial Intelligence: The Past, Present, and Future of AI (pp. 7–8). ABC-CLIO.

United States Department of Health & Human Services. (2021). Trustworthy AI (TAI) Playbook.

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology*, 34(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Wadsworth, B. J. (2003). *Piaget's Theory of Cognitive and Affective Development: Foundations of Constructivism* (5th edition). Pearson College Div.

Wu, K. (2019). *An Empirical Study of Blockchain-based Decentralized Applications* (arXiv:1902.04969). arXiv. https://doi.org/10.48550/arXiv.1902.04969

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A Survey of Human-in-the-loop for Machine Learning. *Future Generation Computer Systems*, 135, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Yang, L., Li, J., Elisa, N., Prickett, T., & Chao, F. (2019). Towards Big data Governance in Cybersecurity. *ata-Enabled Discovery and Applications*, 3(1), 10. https://doi.org/10.1007/s41688-019-0034-9

Yapo, A., & Weiss, J. (2018). Ethical Implications of Bias in Machine Learning. Hawaii International Conference on System Sciences 2018 (HICSS-51).

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., Kararigas, G., Kringen, P., McCullough, M., Möslein, F., Mushtaq, N., Roig, G., Stürtz, N., Tolle, K., Tithi, J. J., van Halem, I., & Westerlund, M. (2021). Z-Inspection®: A Process to Assess Trustworthy AI. *IEEE Transactions on Technology and Society*, 2(2), 83–97. https://doi.org/10.1109/TTS.2021.3066209

Zhang, P., Ding, S., & Zhao, Q. (2023). Exploiting Blockchain to Make AI Trustworthy: A Software Development Lifecycle View. *ACM Computing Surveys*. https://doi.org/10.1145/3614424

Zou, W., Lo, D., Kochhar, P. S., Le, X.-B. D., Xia, X., Feng, Y., Chen, Z., & Xu, B. (2021). Smart Contract Development: Challenges and Opportunities. IEEE Transactions on Software Engineering, 47(10), 2084–2106. *IEEE Transactions on Software Engineering*. https://doi.org/10.1109/TSE.2019.2942301