Towards Blockchain-based AI Alignment Using a Proof of Personhood Consensus Mechanism

Abstract—**Recent developments in the field of AI raised the possibility of timely emergence of a human-level artificial intelligence, and thus increased the urgency of AI safety. As AI alignment lags behind other developments in AI, this work explores how blockchain technology can support AI safety and how society can be involved in AI alignment. Other works propose the use of blockchain technology for the purpose of AI safety, but mostly without referring to novel consensus mechanisms, providing theoretical formulations or describing the potential societal impact of their implementation. This paper suggests to write AI alignment rules in a blockchain that can only be updated by humans which is ensured by the Proof of Personhood consensus mechanism, provides a theoretical formulation of this approach and discusses its significance to society. Thus, relevant concepts are identified, combined and theoretically formulated to propose a system that is protected from AI interference and enables society to obtain and retain control over AI. The creation of such a system is of importance in an era where a human-level artificial intelligence could emerge that may one day reject human ethics, goals, and principles.**

Index Terms—**Artificial Intelligence, AI Alignment, AI Safety, AI Shield, Blockchain, Proof of Personhood**

I. Introduction

Steadily increasing computing power and recent advancements in the field of artificial intelligence (AI) [1] indicate that our generation may witness a human-level AI [2] that could be superior to humankind [3]. This would allow AI to perform self-improving operations in the foreseeable future [4], [5], which may lead to an uncontrolled intelligence explosion threatening humanity [3]. In grim scenarios, a human-level AI could regard the latter as a competitor for resources and could, therefore, try to outperform humans economically, or could even come to the most unaligned conclusion: to be better off without their creators [6]. Due to its superior intelligence, increased adoption and pervasion – imagine billions of interconnected AI-enriched smart devices in future IoT scenarios [7] – this web of AI-controlled machines might find ways to achieve its unaligned goals, without humans to realize and prevent this attempt in time [8]. This constitutes a major threat for humanity [9].

Literature on AI alignment and AI safety [6], [8], [10], [11] proposes ways to create a *friendly* human-level AI [12] and points out difficulties in this regard [13], [14]. Several works highlight advantages of using blockchain technology for AI alignment [6], [8], [15]. However, blockchain-based AI alignment also comes with disadvantages [8], for instance, AI possibly attempting to gain a large percentage of blockchain-based tokens, resulting in a disproportionately large voting power in the system [15]. As AI development accelerates tremendously, it is more important than ever to implement robust AI alignment [10]. Currently, AI safety mainly depends on the AI developers and the ML validation techniques used in their AI models. However, in the ongoing AI arms race, AI alignment may fall short in practice. The aim of this paper is to propose a high-level framework of blockchain-based AI alignment, identify relevant concepts, make improvement suggestions for existing approaches and discuss how society can be more involved in AI alignment. Therefore, this study addresses the following research question (RQ):

- RQ1: **How can blockchain-based AI alignment be improved?**
- RQ2: **How can the proposed AI alignment approach be theoretically formulated?**
- RQ3: **How can the proposed AI alignment approach add value to society?**

The purpose of this study is to propose a new conceptual model of how AI alignment rules can be stored in a tamper-proof and transparent way. It also discusses how to ensure that rules which are implemented in AI shields can only be updated by humans. For this purpose, this work suggests a blockchain-based AI shield using a Proof of Personhood (PoP) consensus mechanism. Hence, formal AI alignment rules stored in a blockchain are used for verifying AI models and their output. In order to achieve this goal, an integrative review [16] was conducted to combine different concepts and perspectives from relevant fields of research.

The structure of this study is the following: first, the stages of machine learning (ML) system development together with selected AI model validation techniques, and the concepts of blockchain and AI alignment are outlined (Chapter II). Then a concept of blockchain-based AI alignment using PoP is presented that aims to be secure and privacy-preserving (Chapter III). In a next step, the proposed concept is theoretically formulated (Chapter IV). Lastly, the societal impact of the proposed approach is discussed (Chapter V).

II. Background

In this chapter, the basic technologies and concepts relevant to the topic of this study are delineated. At the beginning, two selected relevant concepts of AI are described. This includes a sequential overview of the AI model lifecycle and a discussion of selected AI model validation techniques. Then blockchain technology and differences in permission modes and consensus mechanisms are outlined. Lastly, the importance of AI alignment is pointed out and related concepts such as AI shields are introduced.

A. ML System Development

ML represents a technique that includes data analysis in order to automate the creation of analytical models which incorporate sequential stages. The main aggregated stages in a ML system development are *system design*, *system development* and *system operation* [17]. The stages on which this paper focuses are *data development* and *model development* within system development, *automated monitoring* within system operation and *system testing* within both of these aggregated stages [17]. These stages collectively contribute to building an AI model that can learn and make predictions or decisions based on input data.

The *data development* stage includes the collection, preparation and management of data in order to ensure its quality, as this sets up the basis for the development of a ML model. The *model development* stage involves the creation of an AI model including the training of the model. The training process involves the use of a large amount of training data which serves as input for the ML model. During this process weights stored in multidimensional matrices are iteratively adapted to deliver the most accurate output.

After the training process, in the *system testing* stage, a test data set, which should be new to the model, is used to evaluate how well the model has learned from the training data. Within this process, the ML model generates predictions which are then used to refine the ML model by validating its accuracy in an iterative process. Finally, in the *automated monitoring* stage, the ML model is checked in terms of biases or anomalies that lead to incorrect or unaligned outputs.

From a user's perspective, a direct interaction with an AI is enabled via an encoder tool that transforms user input from a human-readable form into a tokenized machine-readable form [18]. The encoded text is then fed into the model predictor. This allows the ML model to generate a reply with the highest accuracy. The respective reply undergoes the decoding process, which reverts it to human-readable text. The response is then being sent to the end user via the user application, which displays an answer in human readable form. This process is shown in form of a sequence diagram in Fig. 1.

B. AI Model Validation

AI model validation techniques are proposed to increase the correctness of AI responses. In particular, these techniques should ensure reproducible and tamper-resistant AI responses, where the security of the AI model output should be guaranteed regardless of the input [19]. As a result, these techniques are a fundamental part of AI safety within a specific AI agent and its respective ML models. Selected AI model validation techniques are described below and grouped by the AI model lifecycle stages.

In the data development stage, *data versioning* makes it possible to track which model is used, depending on which data set was used to train this model [17]. Furthermore, *data validation* is used in order to ensure that the quality of data is appropriate for the training of a model with respect to a specific task [17]. In the model development stage, *digital signatures* can be used for signing a model to ensure its integrity over time [20]. If the model is altered, the corresponding hash values would differ from the original. Another technique is *model watermarking* [21], [22] where a unique identifier or pattern is embedded within the model during the training process. After attempts to alter the model, the watermark can be used to verify its authenticity. This technique ensures that the watermark remains robust against various attacks [21]. In *federated learning* the ML model is trained across multiple devices or nodes. Training the AI model on different computers prevents a single point of failure and increases tamper-resistance. Decentralization of the learning process increases consistency and integrity of the model [23].

In the automated monitoring stage, *model auditing* is used to perform periodic checks on the model's behavior to ensure it remains consistent with its initial version. Any drift in the model's behavior or predictions might indicate tampering or unintended changes [24]. Furthermore, *continuous monitoring* and *drift detection* ensure permanent control of the model's predictions in real-time and compares them to expected outcomes. Real-time monitoring in detecting and handling drift in ML models is of major importance [25].

ML model validation techniques are used to increase cor-

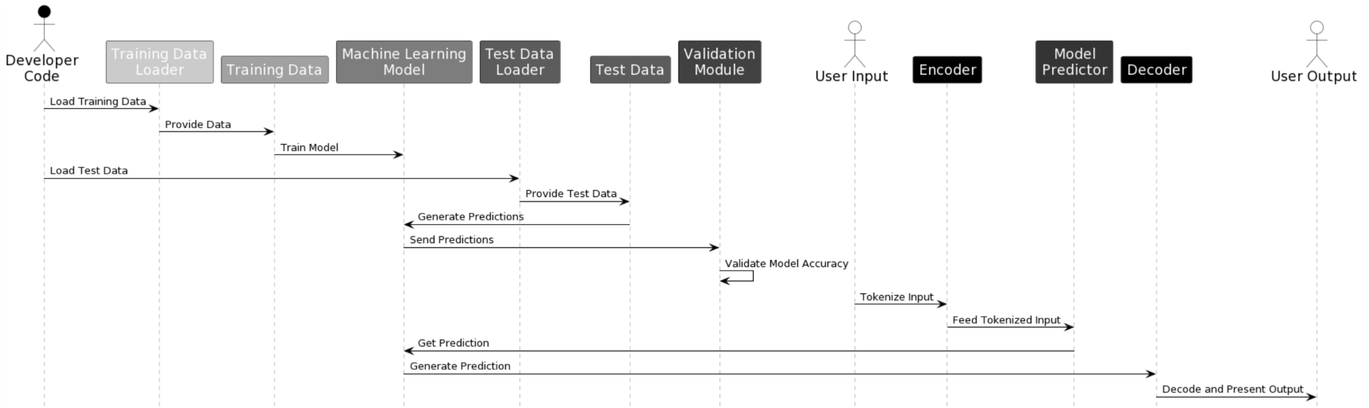

Fig. 1. Simplified user interaction with an AI model.

rectness of AI responses and, thereby, increase AI safety. However, these techniques have drawbacks, especially concerning complex tasks and operating in a general environment [19], [26]. Their deployment depends on the respective AI developers and the availability of validation tools [27], and is not transparent to the public. Hence, additional publicly verifiable and more independent solutions are required to increase safety in AI output.

C. Blockchain

Blockchain technology can be used to address challenges in ML system development. In different stages of the AI model lifecycle, blockchain can be deployed to record the state and usage of a ML model in a decentralized way. Given the transparent nature of blockchain records, any change in the model would be detectable. There are efforts in combining consensus mechanisms from blockchain with ML to ensure model and data integrity [28]. Finally, some of the aforementioned AI model validation techniques, such as digital signatures [20], watermarking or auditing can profit from using blockchain technology.

Blockchains facilitate the transparent and tamper-resistant store of information in a ledger distributed among its participants. For example, the Ethereum blockchain is used to implement a reliable source of truth, for instance, to avoid information manipulation and fake news [29]. Participants agree on a common state of the shared ledger and verify it using a consensus mechanism [30]. The oldest and most famous blockchain, Bitcoin, is governed and secured by the computationally heavy Proof of Work consensus mechanism. Ethereum, the second largest blockchain by market capitalization uses a Proof of Stake consensus mechanism. Other consensus mechanisms required for specific use cases also exist [30], [31]. Additionally, novel alternative consensus mechanisms [30] are emerging, such as Proof of Reputation, Proof of Vote, Proof of Credibility or Proof of Personhood, to name a few.

There are numerous works that propose blockchains to be enriched with AI-based services, for instance, [32], [33], [34]. While proposals for facilitating AI to meet certain limitations of blockchains [33], [35] or blockchain-enabled use cases in the metaverse [32], utilizing AI to meet this goal may pose new challenges, for instance, in regards to AI alignment. In contrast to the works which propose AI to govern blockchain [33], the study at hand proposes a human-governed blockchain to govern AI.

D. AI Alignment

Due to the continuous improvement of AI, its increasing economic potential and impact on our society, the importance of robust AI alignment represents a major factor for the well-being of humanity [10]. Important aspects of AI alignment are verifiability, validity, security and control [10]. The verification and validity aspects of AI alignment face several issues, as it is difficult to formulate and generalize human goals and principles [13], [36], [37], in order to specify them in a way suitable for an evolving AI to interpret them correctly [12], [38], as a human-level AI would operate in a global environment [13], and its operations would have an open and unknown outcome [10]. Since this work focuses on proposing improvements regarding the security and control aspects of AI alignment, verification and validity are out of its scope.

In the context of reinforcement learning, safety throughout all stages of the ML system development is crucial. Reinforcement learning enables an AI agent to adapt its AI model in real time through feedback from its environment to improve achieving its tasks [39]. This type of learning is used for example in robotics [39], where AI agents receive rewards for completing certain tasks or are punished for violating certain predetermined conditions. Thereby, an AI shield can be used to prevent unsafe actions executed by an AI agent.

A current methodology is introduced via proposing a reactive system, which is referred to as an *AI shield*. It acts as a safeguard against actions that could violate predetermined conditions [40]. An AI shield intervenes by monitoring and correcting actions post-decision if they violate predefined conditions or rules. These conditions must be protected against unauthorized changes in order to prevent unsafe actions executed by an AI agent. One way of storing these conditions in a transparent and tamper-proof manner is using blockchain technology. Consensus mechanisms could be used to vote on AI rules allowing society to participate in configuring AI alignment.

III. Towards Blockchain-based AI Alignment Using Proof of Personhood

In this chapter, relevant concepts are identified and combined to answer RQ1. The first step is to identify proposals of blockchain-based AI alignment. In a second step, it is suggested how blockchain technology could be integrated in AI shields. Then it is described how blockchain and biometric features could be used for user authentication which is the foundation for a PoP consensus mechanism. Finally, a high-level overview of blockchain-based AI alignment using a PoP consensus mechanisms is provided.

A. Blockchain-based AI Alignment

AI alignment must be verifiable, valid, secure and controllable [10]. In addition, manipulation resistance is a crucial AI safety criteria and often neglected in AI literature [13]. This paper focuses on the security aspect of AI alignment by proposing AI rules to be stored in a blockchain. Further, it focuses on the control aspect of AI alignment by proposing that updates on AI rules can only be conducted by human agents based on novel consensus mechanisms. Thereby, human agents should also be prevented from carrying out attacks such as Sybil attacks.

Blockchain technology is suggested to improve the robustness of AI alignment by allowing AI rules to be stored in a transparent and fraud-resistant way [8]. Further, the combination of blockchain and AI is proposed to achieve a more robust governance for AI research and development, allow human feedback and establish AI alignment [15]. Hence, humans and blockchain technology can be part of a solution to achieve AI safety [6], [8].

Blockchain-based AI alignment also entails risks [8] as an AI could, for instance, try to accumulate blockchain-native tokens to increase its voting power [15]. This would enable an AI to interfere in the governance model to conduct changes in the AI alignment rules. Established blockchain consensus mechanisms such as Proof of Stake used in the second most popular blockchain, Ethereum, pose this risk. Therefore, the use of alternative consensus mechanisms emphasizing blockchain-based identity [8] must be considered for blockchain-based AI alignment.

Approaches on how blockchain-based consensus mechanisms ensure the integrity of ML validation techniques are being discussed. For instance, a decentralized federated ML is proposed, where safety concerning model training should be guaranteed via incentives [28]. The incentive mechanism is theoretically formulated and should increase model accuracy, data privacy and minimize illicit activities.

B. *Blockchain-based AI Shields*

As mentioned in the context of reinforcement learning, AI shields act as a safeguard against potential violations of predetermined safety and logical rules. In multi-agent reinforcement learning scenarios, AI shielding has been evolved from centralized to decentralized methods [41]. This approach shows that AI agents when equipped with decentralized AI shields perform comparably to those with centralized shields. Thus, shielded reinforcement learning can be applied in settings with decentralized training and execution [41].

Implementing an AI shield on a blockchain could introduce additional layers of transparency and security. AI rules could be embedded within an AI shield, which can be stored on a blockchain. This provides a transparent, immutable, and decentralized method for maintaining and verifying the integrity of the AI's operational guidelines and safety parameters [40], [41]. This solution can prevent unauthorized modifications of the shield rules. Furthermore, it can ensure that any updates or changes to these rules are transparently recorded on the blockchain, while a history of all alterations can be monitored.

This work proposes a system where AI rules encapsulated within an AI shield are stored on a blockchain. Potential benefits are, first, enhanced transparency in the operation and evolution of AI models. Second, a robust, decentralized record of all AI actions and decisions, which can be crucial for accountability and auditability in various application domains. Third, it provides a secure mechanism to ensure the integrity and immutability of AI operational rules. Lastly, it opens up the possibility of involving diverse stakeholders in the governance of AI behavior through consensus mechanisms. Thereby, robust authentication has to ensure that AI agents cannot interfere with the governance of the AI shield, as discussed below.

C. *Blockchain-based Biometric Authentication*

AI agents must not be able to participate in voting on AI rules to hinder AI interference and ensure that humans retain control over AI. Therefore, as a first step of the PoP consensus mechanism, authentication is required to separate human from AI agents. The importance of robust authentication methods is highlighted [42]. Using biometric features for authentication can have advantages to other methods [43]. For instance, authentication using fingerprints are considered more secure and convenient [44]. Further, the combination of multi-factor authentication and blockchain technology is proposed to increase the security level of authentication methods [42]. Besides that, also auditability can be increased by combining biometric authentication systems with blockchain technology [45]. Finally, privacy is identified as a current issue concerning authentication methods [42].

Biometric authentication systems require the storage of sensitive information, such as the fingerprints of individual human agents. Sensitive data could be collected using, for instance, the InterPlanetary File System (IPFS) to store encrypted fingerprints in a secure and data-preserving way [46]. To further increase security and privacy of the authentication method, sensitive data could be distributed among different blockchain nodes as a shared secret [47]. In addition, the usage of blockchain-based decentralized identity (DID) is proposed [48]. Currently, some proposed systems may not be sufficiently resistant against Sybil attacks [47]. Further concepts towards privacy-preserving and secure authentication and biometric features are proposed by using a key derivation function [49], two-factor authentication [50] and a private key generation mechanism [51].

D. *Proof of Personhood*

Once it is guaranteed that only unique human agents can participate in the system, flexible voting solutions should allow for large-scale participation on deciding upon AI rules. Blockchain-based biometric authentication could be applied on a large scale using smartphones, which are widely adopted globally, for example, using their camera for facial recognition [48]. Thereby, data stored at the service provider should be reduced to a necessary minimum required for confirmation of user data. Authentication methods using smartphones are proposed to enable secure and privacy-preserving blockchain-based voting systems [52]. Finally, this method may be the basis for identity-based consensus mechanisms [42].

Such mechanisms are necessary to find agreements on governing AI alignment rules stored in a blockchain. Complementary to biometrics, well-established e-governance services could also be part of the PoP consensus mechanism. For data protection reasons, at least temporarily, they could be used for the authentication of individual human agents until the blockchain-based biometric authentication discussed earlier can guarantee privacy in practice. As most literature on blockchain-based biometrics does not distinguish between authentication and verification, this work suggests using biometric features for verifying human agents and using e-governance services for authenticating them.

E-governance services improve administration and political decision making by facilitating new technologies [53]. These services are based on a common access point which facilitates large scale participation in democratic and more transparent processes. Hence, e-governance can promote democracy [53]. The provision of a sound technical infrastructure and the adoption rate are two critical factors for the success of e-governance services [54]. Therefore, awareness of the need for the proposed AI alignment approach should be established and rewarding participation considered. In case of sufficient user adoption, the discussed mechanisms could be used to govern AI alignment rules. PoP enables society to democratically decide on AI rules which are stored in a blockchain and eventually implemented in an AI shield to specify aligned output of AI agents. Interference of AI agents is prevented by incorporating biometric features and can be, at least temporarily, complemented by e-governance services for authentication to prevent privacy issues.

E. *High-level Overview of Proposed AI Alignment Approach*

In Fig. 2 the proposed high-level schema of blockchain-based AI alignment using a PoP consensus mechanism is shown. Humans can propose goals, ethics and principles, and vote on implementing them in AI rules. However, this requires a successful PoP that uses biometric features and e-governance services to ensure that only humans can participate in these processes. Subsequently, the agreed AI rules are translated into machine code and stored in a blockchain governed by a PoP consensus mechanism. This is to prevent a prospective human-level AI from manipulating data, making proposals and voting on AI rules. Consequently, the AI would fail the PoP and its attempts at altering AI rules. To ensure the security and control aspects of AI alignment, a human-level AI is monitored using the AI rules stored in the tamper-proof blockchain. That helps maintain AI safety and ensures that humans can reap the benefits of AI. This approach could address further AI safety issues [14], such as authenticity & obfuscation, malicious use, manipulation, privacy & integrity, scalable supervision, trust, and transparency & accountability.

IV. Theoretical Formulation of Proposed AI Alignment Approach

By answering RQ2, this chapter provides the basis of the proposed PoP consensus mechanism to ensure AI alignment and the definition of system-relevant parameters through a theoretical formulation. The aim of this section is to provide a theoretical framework to describe the PoP consensus mechanism in a way that different scenarios can be modeled. First, the key variables of the system are defined for this purpose. Second, to draw a picture of the system, several functions are defined to represent the main actions of the agents within the system including influence of those actions on the system.

A. *Definition of Variables*

- N = Set of agents
- P = Set of agents with PoP (real human agents)
- p = Total number of agents with PoP within P
- I = Set of unique identities
- AI = Set of AI agents
- S_P = Set of human agents (P) performing a Sybil attack
- S_{AI} = Set of AI agents performing a Sybil attack
- S = Set of all Sybil attacks
- s = Total number of agents performing a Sybil attack
- f = Function that maps agents to unique identities
- g = Function that detects Sybil attacks
- v = Individual vote of an agent
- W_v = Weight of each vote
- D = Decision

Given that $N = P \cup AI$ and $AI = N \setminus P$, the number of Sybil attacks can be calculated with s being the cardinality of the set S, meaning $s = |S|$, while $S = S_P \cup S_{AI}$. The goal of a PoP system is to ensure that each unique human agent has exactly one identity, which can be represented via a bijective function that maps each agent exactly to one unique identity. This can be defined with the following function:

$$f : P \to I \tag{1}$$

For every $x \in P$, there exists a unique $y \in I$ that $f(x) = y$. Now let $P \subseteq N$ be the set of all agents with PoP that successfully verify as human agents. In order to reach the goal of a bijective relationship between the sets P and I, a function g is introduced to detect Sybil attacks within the set P. With $g : P \cup I \to \{0, x\}$ the function can be defined as:

$$g(x) = \begin{cases} x & \text{if } \forall x \in P, \exists! y \in I \text{ such that } f(x) = y \text{ and} \\ & \forall y \in I, \exists! x \in P \text{ such that } f^{-1}(y) = x, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

This function outputs the agent x if every agent $x \in P$ maps to a unique identity $y \in I$ and conversely every identity $y \in I$ maps only one agent $x \in P$. If this condition is not met, the output is 0. Concerning the voting mechanism within a PoP system, only agents with PoP are able to vote. With p being the cardinality of the set P, meaning $p = |P|$, the decision, D, can be represented as the sum of all votes within P based on the weight of each vote W_v:

$$W_v = \frac{1}{p} \tag{3}$$

$$D = \sum_{i=1}^{p} W_v \cdot v \tag{4}$$

This formulation introduces the concept of ensuring a bijective property of mapping agents to identities. First the set of all agents N is reduced by the set of AI agents AI via the PoP mechanism resulting in the set of PoP approved agents P. This set then is checked in terms of a bijective property to the set of identities I. Thanks to function g that detects Sybil attacks,

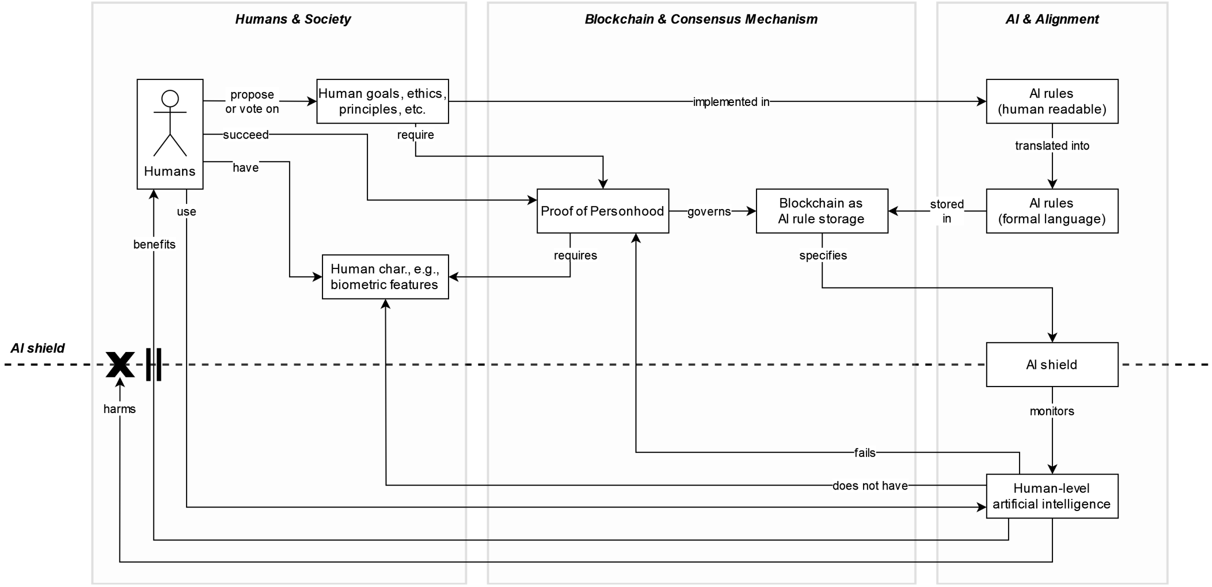
Humans & Society
Blockchain & Consensus Mechanism
AI & Alignment
Humans
propose or vote on
Human goals, ethics, principles, etc.
implemented in
AI rules (human readable)
translated into
succeed
require
have
Proof of Personhood
governs
Blockchain as AI rule storage
stored in
AI rules (formal language)
use
benefits
requires
specifies
Human char., e.g., biometric features
AI shield
AI shield
harms
monitors
fails
does not have
Human-level artificial intelligence

Fig. 2. Proposed high-level schema of blockchain-based AI alignment using a Proof of Personhood consensus mechanism.

this property can be maintained. At this point it is crucial to note that this concept represents an idealized PoP system that distinguishes between AI agents and human agents with 100% accuracy.

In order to establish a reward structure that incentivizes agents to vote correctly, an identity token can be introduced. This identity token could include taxation benefits for all holders of the token. Individuals prove their unique human identity and receive a non-transferable identity token through the PoP mechanism. This PoP-based identity token, including the transactions carried out under its use in the context of governance processes, could be linked to the taxation profile of the human agent that holds the token. In this way token holders could benefit in terms of predetermined tax reductions, either via direct deduction or percentage cuts as long as the individual uses and holds the PoP-based identity token. The tax reduction can be calculated based on the frequency of the token's use. As a result, the more a token holder votes honestly using PoP consensus mechanisms, the greater the potential tax deduction. Current research also shows that enhancement of resilience is addressed to collaborative efforts, which are of major importance in participatory resilience [55]. In this context, there is a general agreement that the higher the number of participants in the system, the higher its resilience [55]. In the context of a PoP system, this means that taxation benefits can contribute to a higher participation in the system, which in return leads to a higher resilience of the overall system. However, a detailed discussion of the implementation of an incentive model in such a system is out of scope of this paper.

B. Identity Verification

In order to check if an agent can be identified as a human agent (verification) and to check if this human agent participates with just one identity (authentication), the following methods can be combined:

- **Biometrics:** In this approach, unique biological properties are used to verify a human individual in order to separate human from AI agents.
- **Government-Issued ID Verification:** In this approach, the system verifies identities via the use of government-issued documents that are assigned to a single human identity.

In this context, the function α can be defined in order to perform two related but distinct tasks being verification and authentication:

- **Verification (1:N):** During this process it is checked, if an agent's identity (template) is already in the system's set of identities. The specifics of α in the verification process depend on the mechanisms used for verifying the uniqueness of an agent. If the agent's identity is already in the system, the verification fails and the agent cannot create a new identity. If the agent's identity is not in the system, the verification succeeds allowing the agent to create a new identity. This process can be represented as:

$$\alpha_{\text{ver}}(a) = \begin{cases} \text{success}, & \text{if } i \notin I \\ \text{failure}, & \text{if } i \in I \end{cases} \tag{5}$$

  where $\alpha_{\text{ver}}(a)$ is the verification process for agent a.
- **Authentication (1:1):** During this process it is checked if a claimed identity matches the agent's real identity. If so, the claimed identity matches the agent's real identity, the authentication succeeds and the agent is authenticated. If the claimed identity does not match the agent's real identity, the authentication fails. This process can be

represented as:

$$\alpha_{\text{auth}}(a, i') = \begin{cases} \text{success}, & \text{if } i = i' \\ \text{failure}, & \text{if } i \neq i' \end{cases} \tag{6}$$

where $\alpha_{\text{auth}}(a, i')$ is the authentication process for agent a with claimed identity i'. In this modification, $\alpha_{\text{ver}}(a)$ represents the verification process and $\alpha_{\text{auth}}(a, i')$ represents the authentication process. The function g, as described above, is mainly involved in the detection of Sybil attacks to identify cases where the same identity is mapped to multiple claimed identities.

C. Sybil Attack Detection

The function g represents the process of identifying Sybil attacks in the system. The specifics of g depend on the strategies used to detect such attacks. Three commonly used methods are briefly described for this purpose:

- **Graph analysis:** In systems that rely on social connections to establish identities, g could be represented as g : Social graph → Potential Sybil nodes.
- **Behavioral analysis:** In this case, g could involve heuristics or ML models, represented as g : User behavior data → Potential Sybil nodes.
- **Periodic re-verification:** If the system periodically requires users to re-verify their identity, g could involve a time component and be represented as g : Verification history → Potential Sybil nodes.

The design of α and g should aim for a balance between effective Sybil attack prevention and maintaining user privacy and system usability.

The proposed PoP-based AI alignment approach was formally described. This builds a precise groundwork for future iterations and scenario. One possible scenario could be its use within the society to increase its control over AI. This can be achieved by using the proposed PoP-based voting mechanism for establishing AI rules.

V. Societal Impact of AI Alignment Approaches

By answering RQ3, it is argued in this chapter how the proposed AI alignment approach could contribute to society by defining AI rules in a democratic way. Although ML validation techniques can increase AI safety, they have shortcomings and are mostly a black box. Another safety issue is the difficulty to correctly annotate a vast amount of training data. Hence, AI alignment cannot always be guaranteed within ML system development. For this purpose, this work suggests a model-independent AI alignment approach that involves society, for example, by voting on the rules of an AI shield. This approach allows society to gain and maintain control over AI.

A. AI Alignment as a Black Box in ML System Development

As described in Chapter II, various techniques throughout the stages of ML system development intend to align an AI model. During the data development stage, *data versioning* and *data validation* are commonly used methods in order to track which dataset has been used to train a model and to ensure the quality of the data to be appropriate [17]. During the model development stage, techniques such as *digital signatures* [20], *model watermarking* [21], [22] or *federated learning* [23] can be used to prevent a model from being altered over time. During the automated monitoring stage, *model auditing* represents a technique to ensure that a model remains consistent with its initial version [56]. In addition, techniques such as *continuous monitoring* and *drift detection* are commonly used [25]. Although these methods are justified and seem to work well within their boundaries, some major issues still persist [19]. Essentially, these techniques are a direct part of the ML model of a private AI developer. These ML models are usually a black box that prevents public verification. Consequently, society must trust the private sector to implement effective AI model validation techniques to increase AI alignment.

Despite the numerous AI model validation techniques, there are universal problems that can affect AI safety. In the data development stage, an AI alignment issue related to a common phenomenon in information technologies, which is *garbage in, garbage out*, may apply. Classifying, filtering and limiting the noise in ML model input data by correctly annotating it reduces the AI safety risks [57]. Ideally, ML models could be trained with cautiously selected data, in the sense of *AI aligned data in, AI aligned data out*. However, there is a trade-off between high variety in input data and the threat of thereby feeding the AI with unaligned data [58]. For general purpose AI models, a vast amount of data must be used to cover the widest spectrum of knowledge possible. In this respect, there is intense competition between AI developers, where a larger training data set is seen as advantageous. The more extensive the training data set is, however, the more difficult it becomes to guarantee safety from a data perspective. Therefore, the competitive pressure between AI developers could threaten AI alignment and thus harm society.

In the model development phase, the problem of hallucinations can occur, which also impairs AI alignment. As generative AI sometimes hallucinates in terms of accuracy, for instance, chatbots which generate answers that appear credible but make no sense in terms of content [57], it may also hallucinate in terms of alignment and could thus harm users. Hence, also the best ML model could make mistakes and unintentionally regard an unaligned response as aligned, which limits AI safety [57]. For instance, in chatbot applications the AI generates an answer with the highest accuracy according to the weights in its AI model. However, lack of training data and its incorrect annotation can result in inaccurate answers. Also, a human-level AI may find ways to intentionally hide an unaligned response to circumvent this safety mechanism. Hallucinating AIs can be particularly harmful as incorrect AI responses can appear as a correct answers to the user input.

B. Proposed AI Alignment Approach Involving Society

AI has an increasing influence on society, for instance, supporting economic decision making, trading bots involved in financial markets, autonomous taxis and participants in the supply chain or teaching children with AI chatbots. Although

AI may deliver sufficiently precise outputs to serve these use cases, its decision making process is typically not well understood. As this process is not fully controlled by humans, the output should be governed by humans [59]. Most influential AI models are developed in the private sector, which can lead to AI alignment issues, for example, ML models including their validation techniques being stored as black boxes, unsafe competition between developers or hallucinations, to name a few. The AI arms race between developers and their increasingly powerful ML models can lead to additional AI alignment threats [59]. This further highlights the necessity of a general technical AI alignment approach, as a regulation framework alone cannot solve technical problems, such as hallucinations [59]. The proposed AI alignment approach is independent of the ML model, its developers and implemented in an AI shield to technically restrict the scope of AI agents.

In Fig. 3, the proposed AI alignment approach is shown, which could enable society to obtain and retain control over AI. By deploying the blockchain-based PoP consensus mechanism described in Chapter III and theoretically formulated in Chapter IV, society can have an impact on AI safety by voting on the conditions of the AI shield in a democratic way. These conditions are stored in a transparent and secure way using blockchain technology. Implementing the proposed approach, the resulting AI shield could prevent actions of AI agents which society, consisting of human agents, are perceived unsafe. Furthermore, the proposed approach is independent of private sector AI models. It can ensure safety of last resort in a technical context for multiple AI agents. Democratically elected AI rules are incorporated in the AI shield to control the output of AI agents of the private sector. As a result, human agents can decide on and adjust rules for AI agents stored in a blockchain using a PoP consensus mechanism. This serves two major purposes. Firstly, gaining control to reduce dependence on the private sector, which is currently being discussed at government level and will be regulated, for example, by the European Union in the form of the AI Act. Secondly, retaining control to address future human-level AI that can potentially perform self-improving operations.

VI. Limitations & Future Work

Finally, a non-exhaustive list of limitations within the scope of this proposal and potential future research directions are pointed out:

- How can the PoP consensus mechanism, which is still at an early conceptual stage, be improved, particularly with regard to data protection?
- Could the PoP consensus mechanism also be implemented in ML system development in the private sector?
- How can Sybil attack prevention be improved in the context of AI alignment based on PoP?
- How can AI alignment rules be correctly translated - from human to machine readable form - and formally defined in order to cover the broadest range of possible output?
- How can an a human-level AI be forced to align to AI rules stored in a blockchain?
- How can incentive structures be defined to reward honest participation and prevent fraudulent (human) behavior? In this context, inclusion of the government can be discussed to facilitate a global standardized framework.
- How can technical expert groups and ethical committees participate in defining and formulating AI alignment rules? This would allow humans to vote upon pre-defined rules proposed by experts to reduce margin of error.

VII. Conclusion

In this study, a blockchain-based AI alignment approach using a Proof of Personhood (PoP) consensus mechanism is proposed to increase AI safety in an era of potentially self-improving human-level AI. The concept of a PoP consensus mechanism using biometrics and e-governance services is found to be feasible from a theoretic point of view. The proposed approach allows society to be involved in defining the conditions of a blockchain-based AI shield, which technically restricts the scope of AI agents. This allows human agents to participate in a democratic, transparent and secure AI alignment process. Also, it further enables humans to obtain power over private sector AI and retain power over future human-level AI. As the AI alignment approach proposed in this paper is in an early conceptual phase, further research needs to be conducted, for instance, on the involved methods and incentive structures that facilitate widespread societal participation.

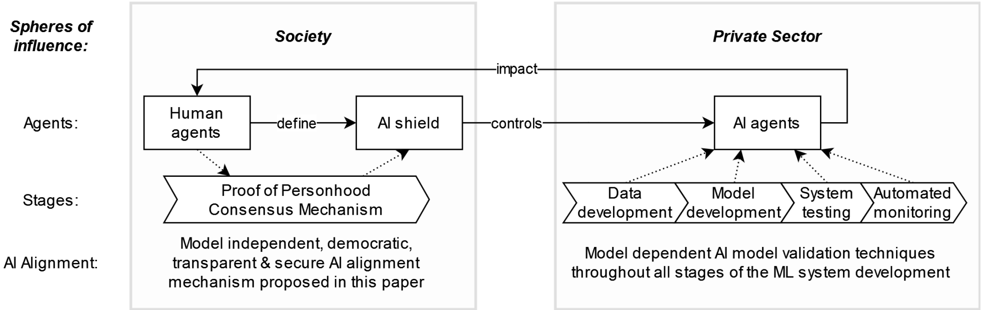
Spheres of influence:
Society
Private Sector
impact
Agents:
Human agents
define
AI shield
controls
AI agents
Stages:
Proof of Personhood Consensus Mechanism
Data development
Model development
System testing
Automated monitoring
AI Alignment:
Model independent, democratic, transparent & secure AI alignment mechanism proposed in this paper
Model dependent AI model validation techniques throughout all stages of the ML system development

Fig. 3. Blockchain-based Proof of Personhood enabling society to obtain and retain control over AI.

References

[1] Lu, Y. (2019). Artificial intelligence: a survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29.

[2] Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., ... & Sowa, J. (2012). Mapping the landscape of human-level artificial general intelligence. AI magazine, 33(1), 25-42.

[3] Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. Fundamental issues of artificial intelligence, 555-572.

[4] Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In Advances in computers (Vol. 6, pp. 31-88). Elsevier.

[5] Yampolskiy, R. V. (2015). Analysis of types of self-improving software. In Artificial General Intelligence: 8th International Conference, AGI 2015, AGI 2015, Berlin, Germany, July 22-25, 2015, Proceedings 8 (pp. 384-393). Springer International Publishing.

[6] Turchin, A., Denkenberger, D., & Green, B. P. (2019). Global solutions vs. local solutions for the AI safety problem. Big Data and Cognitive Computing, 3(1), 16.

[7] Merenda, M., Porcaro, C., & Iero, D. (2020). Edge machine learning for ai-enabled iot devices: A review. Sensors, 20(9), 2533.

[8] Carlson, K. W. (2019). Safe artificial general intelligence via distributed ledger technology. Big Data and Cognitive Computing, 3(3), 40.

[9] Hawking, S., Russell, S., Tegmark, M., & Wilczek, F. (2014). Transcendence looks at the implications of artificial intelligence-but are we taking AI seriously enough. The Independent, 1, 2014.

[10] Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. Ai Magazine, 36(4), 105-114.

[11] Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ... & Legg, S. (2017). AI safety gridworlds. arXiv preprint arXiv:1711.09883.

[12] Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. Global catastrophic risks, 1(303), 184.

[13] Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In Artificial intelligence safety and security (pp. 57-69). Chapman and Hall/CRC.

[14] Hernández-Orallo, J., Martínez-Plumed, F., Avin, S., & Whittlestone, J. (2020). AI paradigms and AI safety: mapping artefacts and techniques to safety issues.

[15] Clifton, C., Blythman, R., & Tulusan, K. (2022). Is Decentralized AI Safer?. arXiv preprint arXiv:2211.05828.

[16] Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of business research, 104, 333-339.

[17] Laato, S., Birkstedt, T., Mäantymäki, M., Minkkinen, M., & Mikkonen, T. (2022, May). AI governance in the system development life cycle: Insights on responsible machine learning engineering. In Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI (pp. 113-123).

[18] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[19] Myllyaho, L., Raatikainen, M., Männistö, T., Mikkonen, T., & Nurminen, J. K. (2021). Systematic literature review of validation methods for AI systems. Journal of Systems and Software, 181, 111050.

[20] Neyigapula, B. S. (2023). Secure AI Model Sharing: A Cryptographic Approach for Encrypted Model Exchange.

[21] Zhang, J., Gu, Z., Jang, J., Wu, H., Stoecklin, M. P., Huang, H., & Molloy, I. (2018, May). Protecting intellectual property of deep neural networks with watermarking. In Proceedings of the 2018 on Asia conference on computer and communications security (pp. 159-172).

[22] Uchida, Y., Nagai, Y., Sakazawa, S., & Satoh, S. I. (2017, June). Embedding watermarks into deep neural networks. In Proceedings of the 2017 ACM on international conference on multimedia retrieval (pp. 269-277).

[23] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

[24] Bayram, F., Ahmed, B. S., & Kassler, A. (2022). From concept drift to model degradation: An overview on performance-aware drift detectors. Knowledge-Based Systems, 245, 108632.

[25] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 1-37.

[26] Kumeno, F. (2019). Sofware engineering challenges for machine learning applications: A literature review. Intelligent Decision Technologies, 13(4), 463-476.

[27] Gao, J., Tao, C., Jie, D., & Lu, S. (2019, April). What is AI software testing? and why. In 2019 IEEE International Conference on Service-Oriented System Engineering (SOSE) (pp. 27-2709). IEEE.

[28] Weng, J., Weng, J., Zhang, J., Li, M., Zhang, Y., & Luo, W. (2019). Deepchain: Auditable and privacy-preserving deep learning with blockchain-based incentive. IEEE Transactions on Dependable and Secure Computing, 18(5), 2438-2455.

[29] Christodoulou, P., & Christodoulou, K. (2020, November). Developing more reliable news sources by utilizing the blockchain technology to combat fake news. In 2020 second international conference on Blockchain computing and applications (BCCA) (pp. 135-139). IEEE.

[30] Lashkari, B., & Musilek, P. (2021). A comprehensive review of blockchain consensus mechanisms. IEEE Access, 9, 43620-43652.

[31] Naz, S., & Lee, S. U. J. (2020, November). Why the new consensus mechanism is needed in blockchain technology?. In 2020 Second International Conference on Blockchain Computing and Applications (BCCA) (pp. 92-99). IEEE.

[32] Badruddoja, S., Dantu, R., He, Y., Thompson, M., Salau, A., & Upadhyay, K. (2022, September). Trusted AI with Blockchain to Empower Metaverse. In 2022 Fourth International Conference on Blockchain Computing and Applications (BCCA) (pp. 237-244). IEEE.

[33] Bouachir, O., Aloqaily, M., Karray, F., & Elsaddik, A. (2022, September). AI-based Blockchain for the Metaverse: Approaches and Challenges. In 2022 Fourth International Conference on Blockchain Computing and Applications (BCCA) (pp. 231-236). IEEE.

[34] Bannazadeh, F., Manohar, M., Mitra, R., Al Ridhawi, I., Abbas, A., & Guizani, M. (2022, September). A Blockchain-enabled and AI-Supported COVID-19 Detection Method. In 2022 Fourth International Conference on Blockchain Computing and Applications (BCCA) (pp. 4-10). IEEE.

[35] Badruddoja, S., Dantu, R., He, Y., Thompson, M., Salau, A., & Upadhyay, K. (2022, September). Making Smart Contracts Predict and Scale. In 2022 Fourth International Conference on Blockchain Computing and Applications (BCCA) (pp. 127-134). IEEE.

[36] Aliman, N. M., & Kester, L. (2019). Requisite variety in ethical utility functions for AI value alignment. arXiv preprint arXiv:1907.00430.

[37] Yudkowsky, E. (2011). Complex value systems in friendly AI. In Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings 4 (pp. 388-393). Springer Berlin Heidelberg.

[38] Russell, S. J. (2010). Artificial intelligence a modern approach. Pearson Education, Inc..

[39] Nagabandi, A., Clavera, I., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., & Finn, C. (2018). Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. arXiv preprint arXiv:1803.11347.

[40] Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., & Topcu, U. (2018, April). Safe reinforcement learning via shielding. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

[41] Melcer, D., Amato, C., & Tripakis, S. (2022). Shield Decentralization for Safe Multi-Agent Reinforcement Learning. Advances in Neural Information Processing Systems, 35, 13367-13379.

[42] Almadani, M. S., Alotaibi, S., Alsobhi, H., Hussain, O. K., & Hussain, F. K. (2023). Blockchain-based multi-factor authentication: A systematic literature review. Internet of Things, 100844.

[43] Ratha, N. K., Connell, J. H., & Bolle, R. M. (2001). Enhancing security and privacy in biometrics-based authentication systems. IBM systems Journal, 40(3), 614-634.

[44] Ogbanufe, O., & Kim, D. J. (2018). Comparing fingerprint-based biometrics authentication versus traditional authentication methods for e-payment. Decision Support Systems, 106, 1-14.

[45] Lee, Y. K., & Jeong, J. (2021). Securing biometric authentication system using blockchain. ICT Express, 7(3), 322-326.

[46] Acquah, M. A., Chen, N., Pan, J. S., Yang, H. M., & Yan, B. (2020). Securing fingerprint template using blockchain and distributed storage system. Symmetry, 12(6), 951.

[47] Sharma, S., Saini, A., & Chaudhury, S. (2023). A Survey on Biometric Cryptosystems and Their Applications. Computers & Security, 103458.

[48] Kim, S., Mun, H. J., & Hong, S. (2022). Multi-Factor Authentication with Randomly Selected Authentication Methods with DID on a Random Terminal. Applied Sciences, 12(5), 2301.

[49] Italis, O., Pierre, S., & Quintero, A. (2023). Privacy-preserving model for biometric-based authentication and Key Derivation Function. Journal of Information Security and Applications, 78, 103624.

[50] Bao, D., & You, L. (2021). Two-factor identity authentication scheme based on blockchain and fuzzy extractor. Soft Computing, 1-13.

[51] Wang, Y., Li, B., Zhang, Y., Wu, J., Liu, G., Li, Y., & Mao, Z. (2023). A novel blockchain's private key generation mechanism based on facial biometrics and physical unclonable function. Journal of Information Security and Applications, 78, 103610.

[52] Abayomi-Zannu, T. P., Odun-Ayo, I. A., & Barka, T. F. (2019, December). A proposed mobile voting framework utilizing blockchain technology and multi-factor authentication. In Journal of Physics: Conference Series (Vol. 1378, No. 3, p. 032104). IOP Publishing.

[53] Von Haldenwang, C. (2004). Electronic government (e-government) and development. The European journal of development research, 16, 417-432.

[54] AlAwadhi, S., & Morris, A. (2009). Factors influencing the adoption of e-government services. J. Softw., 4(6), 584-590.

[55] Sachit Mahajan, Carina I. Hausladen, Javier Argota Sánchez-Vaquerizo, Marcin Korecki, Dirk Helbing, Participatory resilience: Surviving, recovering and improving together, Sustainable Cities and Society, Volume 83, 2022, 103942, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2022.103942.

[56] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[57] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[58] Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. Plos one, 15(12), e0243300.

[59] Kilian, K. A., Ventura, C. J., & Bailey, M. M. (2023). Examining the differential risk from high-level artificial intelligence and the question of control. Futures, 151, 103182.