



# Blockchain for explainable and trustworthy artificial intelligence

Mohamed Nassar<sup>1</sup> | Khaled Salah<sup>2</sup>  | Muhammad Habib ur Rehman<sup>2</sup>  |  
 Davor Svetinovic<sup>2</sup> 

<sup>1</sup>Department of Computer Science,  
 American University of Beirut, Beirut,  
 Lebanon

<sup>2</sup>Center for Cyber Physical Systems,  
 Electrical Engineering and Computer  
 Science, Khalifa University of Science and  
 Technology, Abu Dhabi, UAE

## Correspondence

Davor Svetinovic, Center for Cyber Physical  
 Systems, Khalifa University of Science and  
 Technology, Abu Dhabi, UAE.  
 Email: davor.svetinovic@ku.ac.ae

## Abstract

The increasing computational power and proliferation of big data are now empowering Artificial Intelligence (AI) to achieve massive adoption and applicability in many fields. The lack of explanation when it comes to the decisions made by today's AI algorithms is a major drawback in critical decision-making systems. For example, deep learning does not offer control or reasoning over its internal processes or outputs. More importantly, current black-box AI implementations are subject to bias and adversarial attacks that may poison the learning or the inference processes. Explainable AI (XAI) is a new trend of AI algorithms that provide explanations of their AI decisions. In this paper, we propose a framework for achieving a more trustworthy and XAI by leveraging features of blockchain, smart contracts, trusted oracles, and decentralized storage. We specify a framework for complex AI systems in which the decision outcomes are reached based on decentralized consensus of multiple AI and XAI predictors. The paper discusses how our proposed framework can be utilized in key application areas with practical use cases.

This article is categorized under:

Technologies > Machine Learning

Technologies > Computer Architectures for Data Mining

Fundamental Concepts of Data and Knowledge > Key Design Issues in Data Mining

## KEYWORDS

Blockchain, Consensus, Prediction models, Reputation, Smart contract

## 1 | INTRODUCTION

Artificial Intelligence (AI) is currently showing transformative impact on a number of fields and industries. Over 70 years of research and development has resulted in complex deployed AI systems that are starting to face complicated external issues ranging from security to ethics (Awad et al., 2018; Frank, Wang, Cebrian, & Rahwan, 2019). In particular, AI systems are facing two major limitations: susceptibility to biases and adversarial attacks (Chen et al., 2019; Ng, 2019). The susceptibility to biases increases when AI systems correctly perform decisions on a subset of data but they do not perform well across the whole population. While statistical in nature, these biases may entail societal biases at the detriment of a certain community. Embedding AI in socioinstitutional mechanisms is challenging at many levels as discussed in Sileno, Boer, and van

Engers (2018). From another side, the adversarial attacks on AI systems occur when some malicious actors (e.g., a software, or a person) try to manipulate the data which results in wrong decisions (such as misclassification or bad clusters). In addition to this, current AI implementations represent black-box solutions, therefore, and as such, the need for trustworthy explanations about potential biases and adversarial attacks is increasing rapidly.

Furthermore, the lack of explanation regarding the internal data representations and the decisions made by AI systems is currently one of the most important challenges in making AI systems even more widely accepted, especially in mission-critical domains. For example, we tend to blindly accept a decision recommended by a deep learning system in domains such as object recognition, game playing, and chatbots, therefore tolerating the possibility of a wrong prediction or a false positive, but this blind acceptance should not be tolerable when it comes to critical decision-making systems such as security, healthcare, or finance, where human lives or significant assets are at stake. In response, significant efforts are being made to make deep learning much more trustworthy and controllable by humans, for example, an explainable AI (XAI) initiative was launched by the United State Defense Advanced Research Projects Agency (DARPA) (Gunning, 2016). One of the reasons XAI is required is the increasing number of published adversarial attacks against machine learning and deep neural networks in particular. These attacks come in different flavors such as dataset poisoning, adversarial examples, internal network manipulation, and side-channel attacks (Papernot et al., 2016).

Malicious actors may cause random or targeted misclassifications by manipulating the environment around the system, the data acquisition block (e.g., camera or microphone), or the input samples. The attack can be as simple as adding adversarial noise to the input samples and as stealthy as incrementally shifting the decision boundaries during the training process. A prediction system that lacks comprehensive explanations imposes a take-or-leave response policy. If we know that a system may be subject to adversarial attacks, our trust in its output fades away since no reasoning, proof or explanation is accompanying the output.

According to Thompson (1984), it is more important to trust the people who wrote the software than trust the software itself. But nowadays the AI ecosystem is much more complex with many more stakeholders: the philosopher, the AI researcher, the data scientist, the data provider, the developer, the library author, the hardware manufacturer, the OS provider, and so on. Since it is practically infeasible to build trust relationships with and between all stakeholders, put aside decide liability when things go wrong, we take trustworthiness at a technical level in our scope.

We define technical trustworthiness as the qualitative measure of confidence one can objectively assign to the output of an AI system. To give a concrete example, let us consider the set of binary classifiers for images of dogs and cats. A system with 98% training accuracy and 97% testing accuracy is more trustworthy than a system with 100% training accuracy and 90% testing accuracy. A system which explains its decisions based on features of heads and tails is more trustworthy than a system that predicts “dog” whenever the background is just white. The latter is clearly due to overfitting to the training set. To increase its trustworthiness, an AI system must structurally have elements of explainability, consensus, and historical robustness reputation.

A broader definition of trustworthiness, which is based on ethics and morality, is required much more than ever (Dignum, 2017). Nevertheless, technical AI and XAI trustworthiness enables the implementation of moral and philosophical algorithms, that is, algorithms that are based on ethical principles and philosophical solutions of the so-called “trolley problems” (Keeling, 2019).

In this paper, we specify a framework based on the premise that the critical decisions in complex AI systems must be subject to a consensus among distributed AI and XAI agents or predictors hosted by trusted oracles with the assumption that the majority of these agents are honest. Trustworthy AI requirements for resilience to biases and adversarial attacks can be fulfilled to a large extent by blockchain technologies (Salah, Rehman, Nizamuddin, & Al-Fuqaha, 2019). Blockchain can provide the following key features for an XAI system (Suliman, Husain, Abououf, Alblooshi, & Salah, 2018):

- **Transparency and visibility:** All the transactions are stored in a publicly auditable, append-only and transparent ledger. The ledger state and logs of transactions and function calls are stored in a secure, tamper-proof, decentralized manner that is accessible by all participating stakeholders.
- **Immutability:** The blockchain ledger is comprised of timestamped blocks, where each block is secured by a cryptographic hash. Each block contains a group of transactions and references the hash of the block preceding it. Any change in one of the blocks will invalidate the entire chain of blocks.
- **Traceability and nonrepudiation:** Each participating node or user in the blockchain must cryptographically sign each transaction or function call; and each signed item must get verified and validated by the mining nodes of the blockchain.

Transactions become a part of the immutable ledger, and transacting users or nodes cannot deny or repudiate invocation of function calls or transactions.

- **Smart contracts:** A smart contract (SC) is code that governs the interactions among different participants and allows for the execution of contractual and business logic in an automated, trusted and decentralized manner, where the execution outcome of the SC code is validated and verified by all mining nodes, and agreed on by the majority of these nodes.

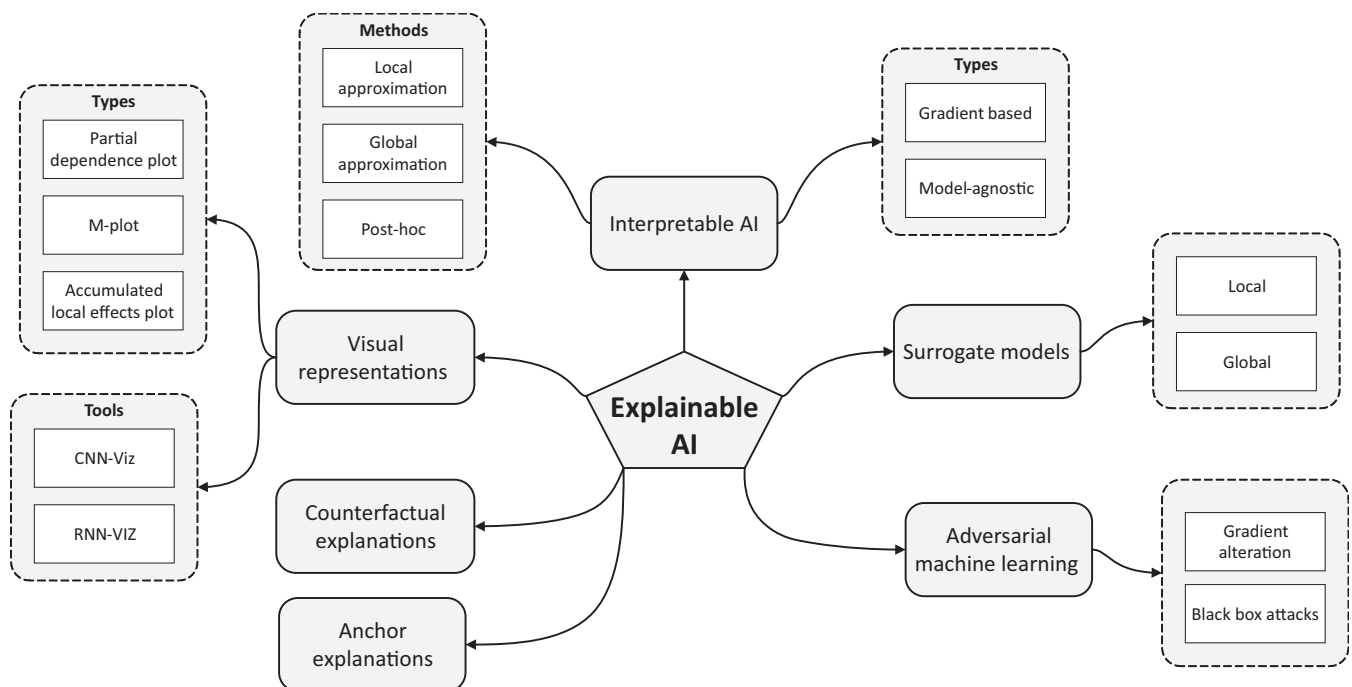
Therefore, a promising solution to address black-box AI is by shifting our trust from a single prediction system to a set of distributed predictors, providing predictions and explanations, in which AI predictions and decision outcomes are recorded, stored, aggregated, and managed in a decentralized, secure, unbiased, and trusted manner by using blockchain, SCs, and decentralized storage.

The contributions of this paper can be summarized as follows:

- We briefly review the state-of-the-art and discuss key issues pertaining to XAI and adversarial AI.
- We specify a general blockchain framework for providing a more trustworthy and XAI and we highlight its key components, functions, and support services that include decentralized registration and reputation.
- We illustrate basic workflow outlining the roles of the four main components used to provide decentralized governance, reputation, and decision-making.
- We discuss different practical use cases to show how our framework can be used.

## 2 | STATE-OF-THE-ART

The European Union's new General Data Protection Regulation (GDPR) presents many implications for AI-based decision systems (Regulation, 2016). Among others, it requires the right of explanation. In case a person is subject to a decision such as automatic refusal of an online credit application or e-recruiting practices, she has the right to obtain an explanation of the decision reached after such assessment and to challenge the decision. GDPR may highlight opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation. Therefore, it is necessary to be able to understand the behavior of AI systems and to limit adversarial situations. Based on related works, we suggest an initial taxonomy of XAI methods as shown in Figure 1.



**FIGURE 1** Types and methods of explainable AI (XAI) techniques



## 2.1 | Interpretable AI

Many machine learning models are inherently interpretable if the training features are primarily meaningful (e.g., the surface of an apartment or the age of an employee). Humans can understand the cause of decisions made by shallow decision trees and sparse linear models since these models can be easily translated into if-else rules. Still, the most successful learning models nowadays are not interpretable since they are mostly based on ensembles of voting submodels such as random forests or are very complex such as boosted trees and deep neural networks. Several interpretation methods are proposed to deal with these cases.

Most of these methods are post-hoc (i.e., they consider a trained classification or regression model), and a few are intrinsic (i.e., they are trained to solve a decision task and provide explanations at the same time). Some methods consider global or modular approximations of the learned model while others provide explanations on a local scale for a single prediction or a group of predictions. The interpretation methods can also be divided into two main approaches: gradient-based approaches which are specific to neural networks, and more general, model-agnostic approaches based on local perturbations and counterfactual examples (Robnik-Sikonja & Bohanec, 2018).

Model-agnostic methods treat the machine learning models as black-box functions (Ribeiro, Singh, & Guestrin, 2016a). XAI also considers the question of devising viable criteria for evaluating the quality of explanations. Even though the usual consumer of explanations is the human end-user, fidelity metrics are required when it is difficult to have a human in the loop to judge good from bad explanations. This is particularly important when the same prediction is given inconsistent explanations, which is also known as the *Rashomon effect*.

## 2.2 | Visual interpretation

Visual interpretation makes use of plots to show the importance of features with respect to the final decisions of the model. The partial dependence plot (PDP) works by marginalizing the machine learning model output to show the relationship between the predicted outcome and a subset of the features, usually only one or two. A better alternative is the M-plot which takes conditional probabilities in the dataset in consideration, therefore avoiding uncommon combinations of features. The accumulated local effects (ALE) plot accumulates the differences in prediction for the data points per interval of the examined feature, by predicting for the extremities of the interval instead of predicting for the feature value itself. This cancels the effect of correlation with other features which remain unchanged.

Individual conditional expectation (ICE) and centered ICE plots are similar to PDP but these plots consider the data instances one by one and track the change in prediction with respect to the change in only one feature. Recent advances in visual interpretation for deep learning are discussed in Choo and Liu (2018).

## 2.3 | Feature contribution

Methods that are not necessarily visual are more important in the context where machines have to take the decision rather than humans. Actual metrics for feature importance and feature interaction were also investigated (Datta, Sen, & Zick, 2016). An interesting idea is to consider that features are players in a coalition game, the prediction value as the total reward of this game and to represent the individual feature contribution by its *Shapley value* (Strumbelj & Kononenko, 2014).

## 2.4 | Surrogate models

A model-agnostic approach is to take a machine learning box and try to simulate it using a surrogate model which happens to be interpretable (Ribeiro, Singh, & Guestrin, 2016b). The goal is to find a simple function  $g$  among a set of interpretable functions  $G$  that best emulates the original function  $f$ . For example,  $g$  is a shallow decision tree while  $f$  is a convolutional neural network. While this task seems intractable at the global level, it suddenly becomes viable at a small, local level. LIME is a method to find local surrogate models by minimizing a loss function. It takes an instance  $\mathbf{x}$  and looks for a function  $g$  which is very similar to  $f$  in the neighborhood of  $\mathbf{x}$ . The exact definition of the neighborhood depends on the data type whether it is images, text, or relational data.

## 2.5 | Counterfactual examples

Most important to our context of critical AI-based decision-making is the ability to explain a single decision (Wachter, Mittelstadt, & Russell, 2017). A counterfactual explanation is a statement of how the world would have to be different for a desirable outcome to occur. It describes a causal argument of the form: “if  $A$  has not occurred then  $B$  would not have occurred.”

In particular, we look up the smallest change to the feature values that would flip the outcome of the prediction. For example, if turning color from black to white would change the AI decision for a loan, the counterfactual would be: “If the skin color was not black, then the loan would not be rejected!” Technically, counterfactuals are computed by minimizing a loss function which is composed of the norm of change in the feature vector (e.g.,  $d(\mathbf{x}, \mathbf{x}')$ ) and the norm of difference between the perturbed instance prediction and the targeted prediction (e.g.,  $|f(\mathbf{x}') - y'|$ ).

In contrast, an anchor explanation is the subset of features that are sufficient to anchor a prediction regardless of the values of the other features.

We summarize the XAI techniques based on our taxonomy in Table 1. The limitations of the different methods invite distributed consensus protocols and trust models which we aim to provide through a blockchain framework.

## 2.6 | Adversarial machine learning

Counterfactual explanations can be used as adversarial samples to deceive the system (Biggio & Roli, 2018). It was shown that slightly changing the pixel values of an image can lead a convolutional neural network to make wrong predictions (Szegedy et al., 2013). The process is even simpler if one has access to the gradients of the learning model (Kurakin, Goodfellow, & Bengio, 2016). Changing only one pixel is sometimes sufficient to fool the system (Su, Vargas, & Kouichi, 2017). Using such approaches, the changes remain unperceived to the human eye. Moreover, three-dimensional (3D) printed artifacts are shown to deceive a camera-equipped detection system (Athalye & Sutskever, 2017). An adversarial printable patch that can be stuck next to objects was also designed by researchers (Brown, Mane, Roy, Abadi, & Gilmer, 2017). Black-box attacks that do not require internal knowledge about model or training data were also proposed by numerous researchers (Papernot et al., 2017).

## 2.7 | Integrating Blockchain with AI systems

Blockchain augments decentralized AI systems by enabling an open-source and publicly accessible digital ledger which is distributed among AI agents across peer to peer networks (Nebula AI, 2018). It enables AI agents to collaboratively perform consensus and save new decisions on the blocks which could be traced back and difficult to alter. Blockchain provides transparency and visibility of AI decisions to all participating AI agents on the network hence it becomes difficult for AI agents to alter or refuse the decisions (Hasan & Salah, 2019). In addition, the programmable blockchain platforms enable SCs-based programming models for decentralized AI applications which ensure self-execution of AI agents based on predefined terms and conditions (Marwala & Xing, 2018).

Blockchain provides decentralization, determinism, immutability, data integrity, and resilience against several security attacks on AI agents and their data. Alternately, AI systems are normally orchestrated around centralized computing and data storage infrastructures and these systems need to handle continuously evolving data which results in probabilistic and volatile decision-making. However, the integration of blockchain and conventional AI systems ensures improved data security and

**TABLE 1** Comparative table of the different explainable AI (XAI) methods

| Method type    | Examples                           | Strengths   | Limitations   |
|----------------|------------------------------------|---|---|
| Intrinsic      | Decision tree, regression, rules   | Direct explanation                                  | Limited performance                                 |
| Visual         | PDP, ALE, t-SNE                    | User-friendly plots                                 | Limited scope of features (1 or 2)                  |
| Surrogate      | LIME, SHAP                         | Blackbox. It works for complex models/deep learning | It assumes feature independence. Simple local model |
| Examples       | Counterfactuals, anchors           | Blackbox. One-shot explanation                      | Rashomon effect                                     |
| Features set   | Feature importance, Shapley        | Wide scope of features                              | Computational limits                                |
| Gradient based | Structure occlusion, saliency maps | It works for deep learning and CNN                  | Whitebox, zero gradient problem                     |

Notes: ALE, accumulated local effects; CNN, convolutional neural network; PDP, partial dependence plot.



collective intelligence due to consensus-based decentralized data and decision storage mechanisms (Salah et al., 2019). It also ensures improved trust and high efficiency brought by multiparty/multiagent decision-making systems that follow various consensus protocols. In previous work, we presented a detailed taxonomic discussion of key concepts to enable blockchain for decentralized AI applications. Considering the scope of this article, we will limit this discussion in this paper and we will urge the interested readers to follow (Salah et al., 2019) for further understanding. Considering the limitations in recent XAI-related literature and integration benefits of blockchain technologies, we propose a blockchain-based framework to deal with the subtleties of explainable and adversarial AI decision-making.

### 3 | BLOCKCHAIN-BASED FRAMEWORK FOR TRUSTWORTHY AI

While XAI looks very promising, explaining complex models can eventually lead to a dilemma. Providing an explanation intrinsically means that a simpler model can be devised. In an interview with one of the pioneers of deep learning, Geoffrey Hinton warned regulators that insisting on making AI systems explain how they work would be a complete disaster (Simonte, 2018). In analogy to humans, most people have no idea how they perform and do work. If people are asked to explain their decision, most of them are forced to make up a story. Neural nets are similar, they learn billions of numbers that represent the knowledge they have extracted from the training data and use these numbers to make a prediction, say whether a pedestrian is in an image or not. If there were any simple rules for deciding whether an image contains a pedestrian or not, it would have been a solved problem long time ago. Nevertheless, having many audited systems to decide on the prediction outcome and produce consistent explanations can make such systems more trustworthy.

Our approach follows this direction of polling multiple nodes or predictors that run AI computational models, and provide explanation to AI outcomes, and gradually build a reputation for each of these predictors. In our approach, we allow the systems to provide explanations of their predictions either directly or through other model-agnostic explanation systems. A plausible explanation of a prediction would definitely contribute to the reliability of the decision made. The interpretation of explanation systems, including the decision outcomes, need to be audited in an immutable, tamper-proof, and decentralized way, and in a way that can be traced and tracked with high reliability and resiliency. Resiliency is robustness against tampering attacks. In case a party changes one bit of information in a block, the hash of the block, which must lead to the previous block in the chain, is broken. Subsequently the whole chain of hashes is broken since it will not lead to the genesis block, even when all previous blocks have been tampered with. Reliability comes from the fact that each node has a full copy of the ledger, and hence if any node fails or goes off, the blockchain remains unaffected. These latter features can be facilitated and provided by blockchain platforms and decentralized storage systems.

#### 3.1 | Honesty and incentives

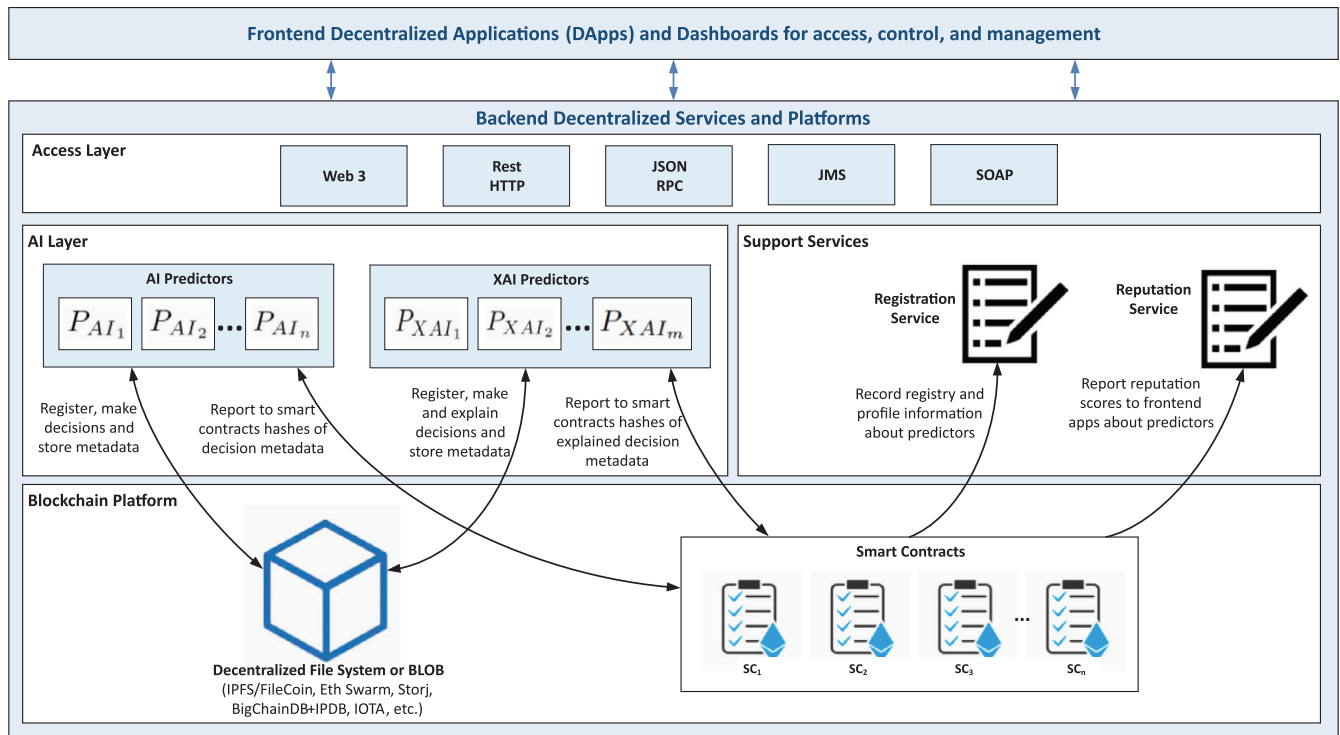
The ecosystem of our proposed solution includes AI and XAI nodes or predictors that act as trusted oracles, perform computation and interact with blockchain SCs which record and log execution outcomes and decisions in the blockchain immutable ledger. The fact that execution and decision outcomes by predictors get validated and verified, and then logged in the ledger in an immutable manner, and accessible by all participants is a strong incentive for an AI system to act honestly in carrying out computation. Moreover, penalties can be imposed in case of reporting false outcomes. SCs can reach final decision outcomes by vote of the majority of the reported outcomes.

In this context, the predictors act as oracles that report results to the SCs. Prior to the use, these oracles have to be registered with public or private owner, and each oracle has a reputation that gets built over time by users through the use of blockchain SCs. Reputation and registration of these oracles are done in a decentralized manner through the use of blockchain SCs, as we discuss later in our proposed framework shown in Figure 2. A good reputation would thrive the business of the oracle owner in the AI market and therefore is a very viable incentive for accurate and correct conduct.

#### 3.2 | Participants

The participants or stakeholders in our proposed solution primary include the following three types:

- Frontend users who are interested in performing a task requiring AI prediction or computation on data (e.g., classification, clustering, regression, etc.). The user may also require explanations about some decisions and provide positive or negative feedback about the decisions.



**FIGURE 2** Blockchain framework for trustworthy artificial intelligence (AI)

- A Predictor ( $P_{AI}$ ) which has a trained AI model for a given task.
- A Predictor with Explanations ( $P_{XAI}$ ) which has a trained AI model and a corresponding model-specific interpretation model.

### 3.3 | Blockchain design and framework

In this section, we present our blockchain-based design and framework to achieve explainable and trustworthy AI. The framework is leveraging blockchain SCs to record, govern interactions, and provide consensus for AI predictions and outcomes among AI and XAI oracles. The framework includes also decentralized storage, registration, and reputation support services. Figure 2 presents the overall system architecture and design of our proposed framework. The framework components are grouped into two distinct subsystems: (1) frontend decentralized applications (DApps) and (2) backend decentralized services and platforms.

#### 3.3.1 | Frontend DApps

Our proposed system for running explainable and trustworthy AI applications supports a variety of decentralized frontend applications that can be used by the different users, stakeholders, or interested parties. The DApps can come in different forms ranging from CLI-based interface applications to interactive mobile or web-based dashboard applications. The interface for the frontend DApps allows configuration, various parameterization, selection of the number of AI or XAI predictors/oracles and their types, accessibility of data, decision outcomes, reputation and registration services, interpretability, and traceability of decision outcomes.

#### 3.3.2 | Backend services and platforms

The system components of our framework *backend* are categorized and grouped into four modules namely: (1) Access layer, (2) AI layer, (3) support services, and (4) blockchain platform.

### *AI access layer*

The *Access Layer* enables interfaces for variety of data transfer protocols. It enables Web3 interface for direct communication between DApps and blockchain platform. It uses JSON-RPC API which facilitates in data transfer between web-enabled applications and Ethereum blockchain network using remote procedure calls (RPC). The JSON-RPC is a light-weight, state-less RPC protocol which enables multiple levels of communication via sockets, processes, http, and a variety of other message passing environments using JSON's RFC 4627 data format. Moreover, the layer includes conventional communication protocols and APIs, such as REST Http to communicate with cloud data centers, JSON-RPC for client-server communications with centralized repositories, Java message service (JMS) API for communication within physical and virtual application components, and simple object message protocol (SOAP) for data communication between sensors-based data sources and backend blockchain platform.

### *AI layer*

AI layer is the primary layer of our proposed framework whereby all data processing and knowledge discovery operations to produce trustworthy, collaborated, and agree-upon decisions are performed. The layer is composed of two different types of predictors; namely: AI predictors ( $P_{AI\_1}$  to  $P_{AI\_n}$ ) and XAI predictors ( $P_{XAI\_1}$  to  $P_{XAI\_m}$ ), as shown in Figure 2. Depending on the configuration received from frontend DApps, AI and XAI nodes either run on raw data and perform all preprocessing operations (such as data cleaning, noise removal, outliers' detection, feature extraction, dimensionality reduction, etc.), or the nodes directly perform decisions on already processed data by inputting into learning models and generating decisions accordingly.

The AI predictors operate on the data using conventional black-box AI algorithms, and produce decisions. On the other hand, the XAI predictors differ in processing mechanisms whereby they produce summary information in addition to decision outcomes to assist in explaining these outcomes. XAI predictors may also communicate with other AI predictors to provide them with explanation capabilities and build a summary of explainable decisions.

It is worth mentioning that there is a cost-trust trade off in selecting the number of predictors. Increasing the number of predictors can obviously increase majority reporting and the mitigation of collusion and dishonesty among predictors, but at the same time, there will be an obvious increase in cost of leasing more prediction nodes to perform the AI task.

### *Support services*

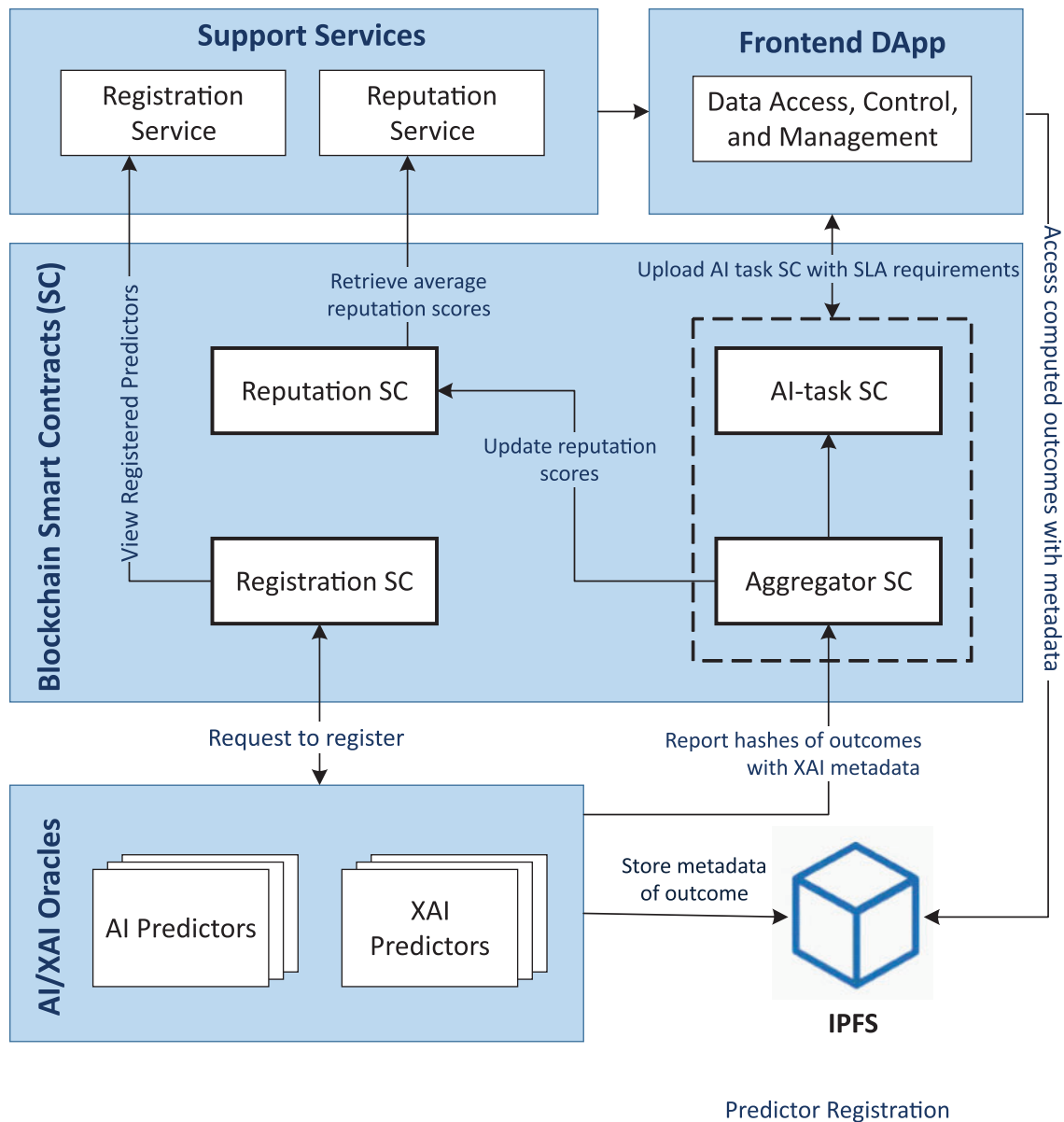
This component provides two types of support services; namely: (1) registration service and (2) reputation service. The registration service enables registering and managing actors and participants in the ecosystem. These participants include users, AI and XAI prediction service providers, data hosting services and repositories, decentralized storage services, as well as reputation services. The reputation service computes and maintains (in a decentralized manner using SCs) the reputation of AI and XAI predictors. Prior to running AI applications, users through DApps can select predictor nodes that are highly reputable. SCs responsible to run AI applications on predictors can be set up to automatically report a reputation score to a reputation SC at the completion of running each AI task. AI and XAI predictors that will not report decision outcomes matching the majority of the predictors will get a negative score, and be penalized in terms of payment rewards which will be paid automatically in cryptocurrency. This forces predictor nodes to act honestly.

### *Blockchain platform*

The **Blockchain Platforms** include primarily: (1) the blockchain network to run different SCs, and (2) a decentralized storage to store results and metadata reported by the AI and XAI predictors. Blockchain is not suitable and highly expensive for storing large size data and content. Also traditional centralized cloud storage or local storage cannot be trusted and can be single point of manipulation, tampering, compromise, and failure. Decentralized storage systems (such as IPFS, Eth SWARM, or Storj, etc.) are stronger and viable alternative. IPFS (interplanetary file system) is a content-addressable, peer-to-peer file system in which the file content is stored on multiple IPFS nodes, and the content hash is the actual address of the file (Benet, 2014). The hashes are used by the SC code for comparison and matching.

The SCs are of different types as depicted in Figure 3. The **Registration SC** enables decentralized registration and identification of AI/XAI predictors on the blockchain network. This SC allows for associating attributes related to ML/AI compute capabilities, performance, URL, and so on. **AI-task SC** is responsible for collecting the final decision outcome from the **Aggregator SC** and reporting the results back to the frontend DApp. The Aggregator SC is responsible for receiving and comparing the outputs (in form of hashes stored on IPFS) from the registered predictors which were previously selected by the user through frontend DApps, and subsequently reporting the final decision to the AI-task SC based on the majority.





**FIGURE 3** Basic operational workflow of the framework

Moreover, the Aggregator SC is responsible for dispersing payments to the oracles as well as reporting a reputation score for each predictor to the **Reputation SC**. Dishonest predictors will be penalized in terms of payment reward and reputation score. Reputation SC will receive scoring from the different Aggregator SCs and compute the temporal accumulative score for predictors.

### 3.4 | Basic workflow

The frontend DApp will interface directly to the backend decentralized platform and support services through standards web and RPC interface protocols that may include JSON-RPC, JMS, SOAP, and so on. The DApp will have the ability to select registered predictors and define the execution SLA (service-level agreement) parameters including latency, pricing, penalties, and so on. The DApp will also upload the AI task and Aggregator SCs which will govern the SLA agreement, and will notify selected predictors of the sources of data that will be used by predictors to carry out computation and analytics, and report their results to the Aggregator SC. At this stage, communication between DApp and predictors is carried out off the chain (i.e., with no transactions being sent to the SCs through blockchain network). If predictors accept the SLA agreement (or AI task/prediction proposal), they send a commitment transaction of acceptance to the Aggregator SC. The predictor can show

seriousness of such commitment by sending a deposit along with commitment transaction in Eth (Ethereum cryptocurrency), as implemented in Hasan and Salah (2018). This Eth deposit can be the same price of computation reward, and it can be lost in case the predictor fails to act honestly or reports results not in agreement with the majority of results reported by other predictors. If the predictor acts honestly, it will get back twice the deposit amount, which consists of one for the reward price and the other for the commitment deposit.

### 3.4.1 | Oracle registration and predictor selection

Our proposed blockchain-based design and framework will enable trusted oracles, which are publicly available prediction service providers, to register through registration SC. The oracle node can host multiple AI and XAI predictors, and provide registry attributes about prediction and decision-making types and capabilities, pricing model, and its abilities to satisfy different SLA objectives in terms of latency and speed, and to handle various input data types, and so on. The frontend DApps will use such attributes, as well as information obtained from registration and reputation services to select the most reputable, affordable, and well performing predictors to carry out the required prediction.

### 3.4.2 | Decision-making

The proposed framework is perceived to be designed as an open public platform, therefore, different oracles can enable different types of AI/XAI predictors to solve the same problem. For example, a few predictors can use tree-based algorithms and others can use neural network based or probabilistic model-based algorithms for classification decisions. However, the variation between decisions among different AI and XAI predictors will be obvious due to multiple reasons such as the computational structure of AI algorithms, types of inputs and outputs, types of decisions, and the quality of training data used for learning models or optimization algorithms. Therefore, we can broadly categorize AI/XAI predictors into following two types (Bohanec, Borstnar, & Robnik-Sikonja, 2017).

- **Deterministic predictors:** Deterministic predictors produce exact decisions which are purely based on the input and training data. These predictors produce Yes/No, True/False, Positive/Negative, Present/Not-present type of decisions. However, only a few predictors such as supervised binary classifiers (e.g., logistic regression, decision trees, support vector machines, etc.) belong to this category.
- **Nondeterministic predictors:** The nondeterministic or probabilistic predictors produce inexact decisions. Most of the existing AI/XAI predictors belong to this category. The inexact predictors perform all types of AI decisions such as classification, clustering, frequent pattern mining, and optimization on input data.

By design, blockchain SCs produce outcomes that have to be exactly the same in order for all blockchain mining nodes to reach consensus. This entails a transformation of inexact and probabilistic decision outcomes to exact and deterministic outcomes. The transformation of decisions through frontend DApp will be a possible option whereby the frontend DApp will specify the interpretation rules (e.g., accuracy ranges or intervals) as part of SLA proposal, and the predictors will produce exact specified value against each range of inexact decisions.

### 3.4.3 | Aggregated decisions

AI/XAI predictors hash and store the metadata of decision outcomes on the IPFS. These decision outcomes include types of decisions, values of evaluation metrics (e.g., levels of accuracy of classifier), confidence values, explanations about decisions, and types of explanations. The aggregator SC will compare the reported hashes from predictors and it will determine the correct decisions based on the majority. The aggregator SC will also ensure that the predictors are satisfying the SLA requirements set by the user DApp, and it will report the final result to AI-task SC which will report back to frontend DApp for final feedback or approval. The AI-task SC will feedback the data and hashes of correct decision outcomes to all participating oracles to produce correct decisions on future input data.

## 4 | REAL-WORLD USE CASES

Our blockchain framework can facilitate and support diverse types of real-world AI systems and applications, in which AI/ML decision outcomes and their explanations can be more trusted and presented in a manner that is decentralized, tampered-proof, and undisputed with traceability and immutable logs that are accessible by all stakeholders. We list a few use cases to illustrate such wide applicability.

### 4.1 | Medical image diagnosis

Recent research works show the emergence of deep medical image analytics for preventive and precision medicines (Lu & Harrison, 2018). Our proposed framework can advance this research by enabling a decentralized disease diagnose system for radiologists who can deploy DApps for consensus-based disease detection. The frontend DApps will input the radiology images whereby AI and XAI predictors from different radiology labs can produce decisions and explanations to help radiologists to reach a more trustworthy, explainable, traceable, and unbiased conclusion. All interested parties including physicians, radiologists, insurance companies, patients, and their care takers can run DApps to perform decentralized predictions and access the same results and explanations.

### 4.2 | Customer profiling

Banking and insurance companies can also benefit from the proposed framework by analyzing massively generated big customer data on social networks and online web portals. Financial institutions perform a background check of potential borrowers or policyholders to minimize the risks of investments. It is envisioned that predictors will be able to determine the credit-worthy customers by detecting anomalies in a customer's previous credit histories, medical profiles, and demographic information. Regulators, financial institutions, credit bureaus, as well as customers, through frontend DApps, will all have access to undisputed, unbiased, trusted decision outcomes and explanations.

### 4.3 | Tax auditing and fraud detection

Recent regulations by Financial Action Task Force and International Monetary Fund are compelling governments to take strict actions against money launderers and tax defaulters in order to stop illegal funding and streamline tax collections (FATF, 2018). Governments can harness citizen's big data and analyze their banking transactions, income sources, and tax-payment histories to curb money-laundering and tax frauds. Hence, governments can benefit from the proposed framework to detect tax evasions and frauds. Our blockchain-based framework allows interested parties including judges, government, attorneys, and citizens to access trustworthy and undisputed decisions and explanations in which money-laundering and tax-evasion patterns are detected.

### 4.4 | Voting and election predictions

The proposed framework can aid government investigation agencies and interested parties to detect and explain fraud in counting votes, or in illegal activities or financing in election campaigns. Different stakeholders will be able to access traceable, trustworthy, explainable, and undisputed prediction outcomes as a result of performing analytics on different aggregated datasets that may include financial records, social media content, emails, and so on.

### 4.5 | Use cases for real-time AI applications

The examples above involve the applicability of our framework for nonreal-time situation, in which decisions are made on nonreal time data. For trustworthy and explainable real-time prediction decisions, which may involve for example a fatal accident by an autonomous vehicle, our framework can be utilized in analyzing and explaining the decisions made by the autonomous vehicle prior to the accident, based on the collected datasets gathered by the different sensing devices deployed on the vehicle. However, it is envisioned that in the future, as blockchain technology matures and overcomes issues related to performance, speed, and scalability, our framework can be leveraged for real-time predictions that are more trustworthy,

decentralized, unbiased, and resilient to adversarial attacks. This realization also entails minimal network latency and higher prediction speed on the part of the predictors.

## 5 | CONCLUSION

Today's regulations increasingly require the interpretation of AI models and explanations of individual predictions. This has led to a new research domain of XAI. At the same time, the deep learning community is skeptic about the real merits of XAI systems, arguing that analogously to the human brain, the best AI systems do not know how to explain their work. In this paper, we have proposed a blockchain-based framework to create a more resilient, trustworthy, XAI system that can reduce biases and adversarial attacks. We presented several use cases to show how blockchain SCs combined with decentralized storage can be leveraged to achieve a trustworthy XAI. Our framework can be used as a reference model to develop a more trustworthy decentralized AI and XAI systems and applications. Technical trustworthiness as a measure of acceptance of decisions by DApp users have requirements of consensus, economic models and incentives for honesty, explainability, and robustness of predictors. In addition, many more infrastructure requirements are needed such as security, privacy, reliability, usability, dependability, performance, and governance. The emerging blockchain technology seems the most adequate, if not the only one, to fulfill these requirements. Still, many challenges must be tackled, the most important ones are minimizing human in the loop for validating explanations and real timeliness for certain applications.

## CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

## AUTHOR CONTRIBUTIONS

Mohamed Nassar: conceptualization, writing-original draft. Khaled Salah: conceptualization, resources, supervision, writing-review, editing. Muhammad Habib Ur Rehman: writing-review and editing-lead. Davor Svetinovic: resources, supervision, editing.

## ORCID

Khaled Salah  <https://orcid.org/0000-0002-2310-2558>

Muhammad Habib ur Rehman  <https://orcid.org/0000-0001-7428-2272>

Davor Svetinovic  <https://orcid.org/0000-0002-3020-9556>

## RELATED WIREs ARTICLES

[Tunnel crack detection using coarse-to-fine region localization and edge detection](#)

[Hierarchical third-order tensor decomposition through inverse difference pyramid based on the three-dimensional Walsh-Hadamard transform with applications in data mining](#)

## REFERENCES

- Athalye, A., & Sutskever, I. (2017). Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397.
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shari, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59–64.
- Benet, J. (2014). IPFS-content addressed, versioned, p2p file system. arXiv preprint arXiv:1407.3561.
- Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317–331.
- Bohanec, M., Borstnar, M. K., & Robnik-Sikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, 71, 416–428.
- Brown, T. B., Mane, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:1712.09665.
- Chen, T., Liu, J., Xiang, Y., Niu, W., Tong, E., & Han, Z. (2019). Adversarial attack and defense in reinforcement learning-from AI security view. *Cybersecurity*, 2(1), 11.
- Choo, J., & Liu, S. (2018). Visual analytics for explainable deep learning. arXiv preprint arXiv:1804.02527.

- Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy* (pp. 598–617), San Jose, CA.
- Dignum, V. (2017). Responsible autonomy. arXiv preprint arXiv:1706.02513.
- FATF. (2018). *FATF fintech & regtech initiatives*. [Online]. Retrieved from [https://www.fatf-gafi.org/fintech-regtech/fatfonfintechregtech/?hf=10&b=0&s=desc\(fatf\\_releasedate\)](https://www.fatf-gafi.org/fintech-regtech/fatfonfintechregtech/?hf=10&b=0&s=desc(fatf_releasedate))
- Frank, M. R., Wang, D., Cebrian, M., & Rahwan, I. (2019). The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence*, 1(2), 79.
- Gunning, T. (2016). *Explainable artificial intelligence (XAI)*. [Online]. Retrieved from <https://www.darpa.mil/program/explainable-artificial-intelligence>
- Hasan, H. R., & Salah, K. (2018). Blockchain-based proof of delivery of physical assets with single and multiple transporters. *IEEE Access*, 6, 46781–46793.
- Hasan, H. R., & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *IEEE Access*, 7, 41596–41606.
- Keeling, G. (2019). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, 1–15. <https://doi.org/10.1007/s11948-019-00096-1>
- Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.
- Lu, L., & Harrison, A. P. (2018, July). Deep medical image computing in preventive and precision medicine. *IEEE Multimedia*, 25(3), 109–113. <https://doi.org/10.1109/MMUL.2018.2875861>
- Marwala, T., & Xing, B. (2018). Blockchain and artificial intelligence. arXiv preprint arXiv:1802.04451.
- Nebula AI, T. (2018). Nebula ai (nbai)decentralized ai blockchain whitepaper. Retrieved from [https://neironix.io/documents/whitepaper/4082/NBAI\\_whitepaper\\_EN.pdf](https://neironix.io/documents/whitepaper/4082/NBAI_whitepaper_EN.pdf)
- Ng, A. (2019). *AI for everyone*. Retrieved from <https://www.coursera.org/learn/ai-for-everyone/home/welcome>
- Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (pp. 506–519), Saarbrücken, Germany.
- Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *Security and privacy (EuroSec&P)*, 2016 *IEEE European symposium on* (pp. 372–387).
- Regulation, G. D. P. (2016). 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. *Official Journal of the European Union (OJ)*, 59(1–88), 294.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SigKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).
- Robnik-Sikonja, M., & Bohanec, M. (2018). Perturbation-based explanations of prediction models. In *Human and machine learning* (pp. 159–175). Cham, Switzerland: Springer.
- Salah, K., Rehman, M. H. U., Nizamuddin, N., & Al-Fuqaha, A. (2019). Blockchain for AI: Review and open research challenges. *IEEE Access*, 7, 10127–10149. <https://doi.org/10.1109/ACCESS.2018.2890507>
- Sileno, G., Boer, A., & van Engers, T. (2018). The role of normware in trustworthy and explainable AI. arXiv preprint arXiv:1812.02471.
- Simonte, T. (2018). *Googles ai guru wants computers to think more like brains*. [On-line]. Retrieved from <https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/>
- Strumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3), 647–665.
- Su, J., Vargas, D. V., & Kouichi, S. (2017). *One pixel attack for fooling deep neural networks*. arXiv preprint arXiv:1710.08864.
- Suliman, A., Husain, Z., Abououf, M., Alblooshi, M., & Salah, K. (2018). Monetization of IoT data using smart contracts. *IET Networks*, 8(1), 32–37.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
- Thompson, K. (1984). Reflections on trusting trust. *Communications of ACM*, 27(8), 761–763.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 2018.

**How to cite this article:** Nassar M, Salah K, ur Rehman MH, Svetinovic D. Blockchain for explainable and trustworthy artificial intelligence. *WIREs Data Mining Knowl Discov*. 2020;10:e1340. <https://doi.org/10.1002/widm.1340>