applied sciences

MDPI

Article

Trust Management for Artificial Intelligence: A Standardization Perspective

Tai-Won Um [1], Jinsul Kim [2,*], Sunhwan Lim [3] and Gyu Myoung Lee [4]

[1] Graduate School of Data Science, Chonnam National University, Gwangju 61186, Korea; stwum@jnu.ac.kr
[2] School of Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, Korea
[3] Intelligent Convergence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon 34129, Korea; shlim@etri.re.kr
[4] Faculty of Engineering and Technology, Liverpool John Moores University, Liverpool L3 3AF, UK; g.m.lee@ljmu.ac.uk
* Correspondence: jsworld@jnu.ac.kr

Abstract: With the continuous increase in the development and use of artificial intelligence systems and applications, problems due to unexpected operations and errors of artificial intelligence systems have emerged. In particular, the importance of trust analysis and management technology for artificial intelligence systems is continuously growing so that users who desire to apply and use artificial intelligence systems can predict and safely use services. This study proposes trust management requirements for artificial intelligence and a trust management framework based on it. Furthermore, we present challenges for standardization so that trust management technology can be applied and spread to actual artificial intelligence systems. In this paper, we aim to stimulate related standardization activities to develop globally acceptable methodology in order to support trust management for artificial intelligence while emphasizing challenges to be addressed in the future from a standardization perspective.

Keywords: trust management; trustworthiness; artificial intelligence; standardization

Citation: Um, T.-W.; Kim, J.; Lim, S.; Lee, G.M. Trust Management for Artificial Intelligence: A Standardization Perspective. *Appl. Sci.* **2022**, *12*, 6022. https://doi.org/10.3390/app12126022

Academic Editor: Peter Gorm Larsen

Received: 16 April 2022
Accepted: 6 June 2022
Published: 14 June 2022

1. Introduction

A large amount of data are being generated and accumulated in various service fields, and artificial intelligence (AI) technology, which has the ability to extract meaningful information and make decisions by analyzing it, is rapidly developing and spreading [1–3].

AI systems build a model framework by installing software and libraries based on the hardware infrastructure, and then preprocess, analyze, and visualize the input data [4]. Developing and operating these AI systems require sophisticated and complex processes which are very difficult to implement. There is continuous research and development to significantly reduce the complexity and operation time while increasing the accuracy of AI systems [5]. In addition, there are many research activities to emphasize trust for giving confidence concerning the predictions processed by AI algorithms in a system [6–8].

In discussing the characteristics of AI technology, classification according to Weak AI and Strong AI is generally used. Weak AI is an artificial intelligence technology that focuses on relatively narrow areas such as image recognition, voice recognition, and data classification, while Strong AI is an artificial intelligence technology that enables machines to intelligently solve various problems with true intelligence and self-recognition [9].

Weak AI is already widely applied and yields remarkable results; however, there are still many controversies relating to the technical potential and social impact of Strong AI.

AI-based systems and applications can sometimes behave in ways the system designer did not foresee. A recent study from Stanford University found that an AI-based facial recognition algorithm discriminated against homosexuals and heterosexuals on dating sites with up to 91% accuracy, raising tricky ethical questions [10]. This study identified that

faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain. As demonstrated in this study, the widespread adoption of AI technologies is strongly required to proactively consider privacy risks with the need for safeguards and regulations to resolve ethical issues such as unintended discrimination.

The AI system is a black box with very large information asymmetry between developers and users and high uncertainty. Therefore, it is vital to forecast the potential risks of AI systems and to evaluate and manage the trust to increase transparency and accountability. However, due to several reasons, evaluating and managing the reliability of AI systems and algorithms are very complex [9].

Many software and hardware components existing in infrastructure, AI models and data processing platforms, applications/services, etc., compose the AI system through complex connections [4]. The reliability of AI systems is largely driven by algorithms and data. Generally, it will be difficult to achieve a reliable and stable AI ecosystem if there is no way to measure, analyze, and verify trust in AI systems.

This study identifies the requirements for trust management of AI systems to prevent unintended damage caused by malfunctions of AI algorithms and data bias. We propose a trust management framework for AI systems and aim to analyze strategies and issues for international standardization based on these requirements.

Following this introduction, Section 2 reviews the standardization trends related to AI trust management, Section 3 analyzes the existing research status and limitations related to AI reliability, Section 4 proposes the AI trust management framework, Section 5 analyzes the international standardization strategy and related issues for the AI trust management framework and suggests new standardization items, and Section 6 concludes by summarizing this study and suggesting future research.

2. Standardization Trend for AI Trust

This section describes the international standardization trends related to AI and trust management, which is being carried out or in progress by international standardization organizations.

Table 1 summarizes Al-related standardization activities in the International Telecommunication Union–Telecommunication Standardization Sector (ITU–T). The ITU–T Study Group (SG) 13 targeting future networks successfully concluded the Focus Group on Machine Learning for Future Networks (FG–ML5G) and published several recommendations as this FG's results. In addition, the Focus Group on Autonomous Networks (FG–AN) newly established under SG13 is currently developing a deliverable on trust in AN. Furthermore, other SGs such as SG16 have many activities to develop standards on AI in terms of applications and services aspects, with the Focus Group on Artificial Intelligence for Health (FG-AI4H).

Table 2 presents standardization activities on trust. Firstly, ITU–T SG13 started to develop a set of recommendations on trust in the context of Information Communication Technology (ICT) infrastructure and services. Based on these recommendations, other recommendations are still developing and there have been strong demands to consider trust in data and AI aspects. For example, FG–DPM successfully developed a deliverable on data trust and started to identify potential work items on trust in data processing, management, and analytics, including the AI/ML. International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) Joint Technical Committee (JTC) 1 standardization subcommittee (SC) 42 is focusing on overall AI technology. Specifically, there are serval technical reports on trust, ethics, and risk management in AI.

Table 1. International standardization trends related to artificial intelligence.

Standard Body	Standard Group	Standard Document	Main Content
ITU–T	SG13 and FG–ML5G	Y.3172	An architectural framework for the application of machine learning in future networks including IMT-2020.
		Y.3173	A method for measuring the intelligence level of future networks including IMT-2020.
		Y.3170	Data processing framework to apply machine learning to future networks including IMT-2020.
		Supplement 55 to Y.3170-series	Machine Learning Use Cases in Future Networks such as IMT-2020.
		Y.ML-IMT2020-NA-RAFR	AI-based resource control and failure recovery automation in future networks including IMT-2020.
		Y.ML-IMT2020-serv-prov	AI-based user-driven network service provisioning in future networks including IMT-2020.
		Y.ML-IMT2020-MP	Machine Learning Marketplace in Future Networks including IMT-2020
		Y.IMT2020-AIICDN-arch	AI integrated cross-domain network structure in future networks including IMT-2020.
	SG16	Y.Sup.AI4IoT	AI role in IoT data management and implementation of AI-based technology for smart cities
		Y.4472	An open data application programming interface for IoT data in smart cities.
	FG–AI4H	FG–AI4H Whitepaper	White Paper for the ITU/WHO Focus Group on AI Health.

Table 2. International standardization trend related to trust.

Standard Body	Standard Group	Standard Document	Main Content
ITU–T	SG13	Y.3051	Basic principles for a trusted environment in ICT infrastructure.
		Y.3052	Trust provisioning overview for ICT infrastructure and services.
		Y.3053	Trust networking framework with trust-centric network domains.
		Y.3054	Trust-based media services framework.
		Y.3057	Trust index for ICT infrastructure and services.
		Y.trust-arch	Functional architecture for trust-based service provisioning.
		Y.3056	An open bootstrap framework that supports trust networking and services for distributed ecosystems.
		Y.3055	Trust-based personal data management platform framework.
	FG–DPM	TR D4.3	Technical enabler overview for trust data.
	SG17	X.5GSec-t	Trust relationship-based security framework in the 5G ecosystem.
ISO/IEC	JTC1 SC42 WG3	TR 24028	Artificial intelligence trust overview.
		TR 24368	Artificial intelligence ethics and social importance concept.
		CD 23894	Artificial intelligence—risk management
		TR 24027	AI systems and AI-based decision-making bias.
		TR 5254	Goals and methods for exploitability of ML models and artificial intelligence systems.

3. Trust Management in AI Technologies

This section first analyzes the basic configuration and operation procedure of the artificial intelligence system and describes the status of trust management technology related to AI.

3.1. Basic Configuration and Operation Procedure of Artificial Intelligence System

AI systems are generally based on high-performance computing, storage, and networking infrastructure, and include AI platforms, applications, and services that are composed of iterative multi-step data processing and AI models [4].

Artificial intelligence refers to infrastructures such as cloud and edge computing in which AI applications and platforms run and include computing servers, high-performance storage, and high-speed networking. A high-performance Central Processing Unit (CPU), Graphics Processing Unit (GPU), or Tensor Processing Unit (TPU) are the main resources for AI processing in a computing server, and a large amount of system memory is essential. In addition, a storage system with sufficient capacity and performance is required to support data diversity and high-speed processing used for AI learning and processing. High bandwidth and low latency between computing nodes and storage are very important in the cloud and edge computing environments, and a high-speed network that supports them must be provided.

In AI, data are used for training and testing a model and for analysis by the trained model. Depending on the type of AI-based service, database, Internet of Things(IoT), web, social network service, mobile app, public data, etc., will be the source of the data. The accuracy of the trained AI model is directly affected by the quality of the data used to train the model, therefore the selection and validation of data from reliable sources is very important.

An AI model that has been trained using the training data is distributed to the target system and used to analyze and infer new input data in real time. Maintenance, such as retraining and the improvement of AI models, can be applied by applying the latest data as needed.

3.2. Trends Related to ICT Trust Management Technology

Trust in software and hardware and data and information constituting the ICT system must be a prerequisite to provide stable and sustainable ICT services [11].

A trusted ICT service environment can promote the emergence of innovative AI application services and the revitalization of the ICT convergence industry by reducing the transaction costs incurred when using ICT services and transacting data and information.

Trust evaluation is a technical approach for expressing trust relationships and consists of measuring and calculating various properties of the entities, and research and development are in progress in the direction of defining trust metrics and corresponding properties [12].

A method for measuring, quantifying, and evaluating trust is needed to recognize the level of trust of any ICT system and compare the level of trust with other systems. ITU–T recommendation Y.3052 describes an "overview of trust provisioning in ICT infrastructures and services" and a trust attribute, a trust indicator, and a trust index were defined as follows [11]:

- Trust attribute: This indicates the characteristics of an entity and is of qualitative and quantitative types including direct and indirect trust. They represent the attributes and capabilities of trusted entities. Qualitative attributes require a quantification process to accumulate quantitative attributes.
- Trust indicator: This is used to calculate the confidence index by combining the qualitative and quantitative attributes of trust. The objective trust indicator represents the ability to quantitatively represent the trustworthiness of an entity. The subjective trust indicator reflects either the subjective or personal attributes of the trusting entity. A confidence indicator is calculated as a measurement instance of confidence because its value changes over time.
- Trust index: This is a composite and relative value that combines several trust indicators into one benchmark measure of the trustworthiness of an entity similar to an ICT development index or stock market index. It is a comprehensive accumulation of

objective and subjective trust indicators for calculation. The trust index evaluates and quantifies the trustworthiness of the trustee.

For trust evaluation, trust-related data must be collected from various sources. The trust attribute is used to calculate a trust indicator based on the collected data. Trust indicators also have self-accumulating properties in subjective or objective properties. The trust index is calculated in a self-accumulating method by combining the objective trust index and the subjective trust index [13].

Norway's University of Agder analyzed various trust factors such as recursiveness, psychological risk, and reputation of data from the software, hardware, device, and service perspective, and a trust framework is also being researched [14].

In terms of security management, for terminal/sensor security, a trusted technology that enables terminal/sensor device authentication to identify and authenticate whether data are transmitted from the correct device when communicating with other devices is being studied. FP7's uTRUSTit project published "Privacy visualization requirements in the Internet of Things" and conducted research on privacy and trust regarding IoT [15].

Identity trust and privacy are fundamental prerequisites for providing trust-based services, and many global companies are developing trust-based solutions for various fields, including cloud Business-to-Business (B2B) platforms. TRUSTe's cloud-based privacy authentication and management system manages privacy authentication during data collection and provides privacy data collaboration with third-party operators. In addition, studies on techniques for adding privacy properties to data and controlling the flow of information based on trusted computing are being actively conducted [16].

3.3. AI Trust Related Trends

Urs Gasser et al. of Harvard University presented an AI governance structure composed of three layers: a technical layer, an ethical layer, and a social and legal layer [9]. In the AI governance structure of Harvard University, the technical layer plays a key role in the governance management of algorithms and data, the ethical layer judges human rights-based ethical standards and principles as key elements of governance, and the social and legal layer mentioned the establishment of a regulatory body for AI systems and the AI regulatory process.

AI system users need to understand the basis of prediction and judgment of AI systems. Operators of AI-based autonomous systems will also need to make decisions and understand why the system behaves so that they can respond to unexpected system behavior.

Explainable AI (XAI) can justify AI-based decisions and improve algorithms by using explanatory models that link explanatory semantic information to record and analyze the learning and prediction processes of algorithms on data [17]. Users will be able to identify weaknesses in AI systems, predict how the system will behave in the future, and correct errors in the system with XAI technology.

Responsible AI refers to AI that complies with social values, moral and ethical considerations, and has three main characteristics: accountability, responsibility, and transparency [18].

3.4. Analysis of Limitations and Requirements of Artificial Intelligence Trust Research

Existing research on the trustworthiness of AI systems is still at a basic stage with only the abstract characteristics that AI systems should have presented, and the specificity of the system requirement level for AI trust is lacking [7,19]. This section describes trust requirements for the predictable and safe use of AI systems:

- Measurement and calculation: It is difficult to derive trust as a generalized formula due to the diversity of AI systems and differences in their intrinsic characteristics. However, quantifying the level of trust in AI systems is important. It should be able to define measurable AI trust metrics and determine the level of trust in AI systems through trust calculations. The level of trust in AI can be measured by classifying it into an objective method that is quantitatively measured such as quality of service (QoS) or

a subjective method that is qualitatively calculated such as quality of experience (QoE). Different AI services and applications may require different trust attributes [20].

- Trust relationship: In addition to the human-to-human trust that has been reviewed in the traditional social domain, the trust relationship between AI-applied systems and people, AI systems and AI systems, etc., should be defined, and trust-based interactions between them should be analyzed [21–23].
- Trust management: In an AI system, trust interacts with all layers, from the upper AI application to the lower physical layer. Therefore, similar to security, trust management technology is required as a separate common layer that covers all vertical layers. Trust management has key functions such as monitoring management, data management, algorithm management, expectations management, and decision management. Trust information about reputation and recommendations, in particular, can be used to support these functions [24].
- Dynamically changing properties: Trust indicator values for AI systems are not kept constant and may fluctuate depending on data and surrounding circumstances; therefore, continuous tracking and management are required [19].
- Constraint environment: Constraints in hardware performance such as CPU/GPU/TPU, memory, and storage constituting the AI system, the types of AI algorithms applied, and restrictions in data collection must be considered.
- Lifecycle management: Human oversight may be required as a safeguard throughout the lifecycle of an AI system, from design, development, launch, use, and disposal. Risk assessment is vital because the autonomous operation and function update of a specific AI system during its lifecycle can have a significant impact on safety [25].

The indicators below are selected according to the AI application service field, and differential weights are applied according to their importance so that they can be used in the trust index of the AI system. The main factors that can be considered as indicators for quantifying the trust level of AI systems are as follows:

- Data quality: The quality of the data set used in the AI system has a decisive effect on training machine learning algorithms and performing classification and decision-making. Feeding malicious data into the system could change the behavior of AI solutions. It should be possible to remove this data before it is applied to training if the collected data are biased. Validation and testing of the data set should be carefully performed before applying to the AI system, and the data supplied to the AI system should be recorded at all times, and audits should be performed in case of future problems [26].
- Non-discrimination: Direct or indirect discrimination based on ethnicity, gender, sexual orientation, or age may lead to exclusion of certain groups. Discrimination in AI systems can occur unintentionally due to data problems such as bias and incompleteness or design errors in AI algorithms. Those who control AI algorithms may seek to achieve unfair or biased results such as by deliberately manipulating data to exclude certain groups of people [27].
- Privacy protection: Digital records of human behavior contain highly sensitive data such as gender, age, religion, sexual orientation, and political views, as well as in terms of preferences. Privacy and data protection must be ensured at all stages of the AI system lifecycle, including any data provided by the user, as well as any information generated about the user in their interactions with the AI system [28].
- Robustness: AI systems must be robust and secure enough to handle errors or inconsistencies in the design, development, execution, deployment, and use phases to respond appropriately to erroneous results.
- Reproducibility: Despite the complexity and sensitivity of the AI system to training and model-building conditions, it should be able to produce consistent results according to the input data in a given situation. Lack of reproducibility can result into unintended discrimination in AI decisions.

- Accuracy: AI systems must ensure accuracy such as the ability to classify data into the correct categories or the ability to make correct predictions or decisions based on data or models.
- Security: Like all software systems, AI systems can contain vulnerabilities that attackers can exploit. When an AI system is attacked such as by hacking or malware, data and system behavior can be altered, causing the system to make different decisions or shut down the system completely. Cyber security management that can quickly remove and manage vulnerabilities in AI systems as soon as they are discovered and prevent infection of malicious codes such as viruses, worms, and ransomware must be applied.
- Explainability: Explainability should be applied so that the mechanisms by which AI systems make decisions can be interpreted, inspected, and reproduced [29].

Trust management must be performed in all lifecycles of analysis, design, development, and use to satisfy the requirements for guaranteeing the reliability of the AI system presented above. Given that AI systems are constantly evolving and operating in a dynamic environment, an ongoing management process is particularly important in achieving trustworthy AI systems.

4. AI Trust Management Framework

In this section, a hierarchical structure-based framework for measuring and managing trust in AI systems is described.

Layered model modularity is one of the main mechanisms for managing complex systems, and various parts of the entire system are modularized and arranged in a parallel hierarchical structure. Modularity is usually designed and implemented in a way that minimizes interdependencies. The trust management structure proposed for the AI system may either be applied to several layers as a whole or on only a specific layer if necessary.

4.1. Trust Target and Management Elements by a Layer of Artificial Intelligence System

In this section, as shown in Figure 1, the AI system is classified into the data layer, model layer, and application layer, and the trust target and trust management elements in each layer are analyzed and described.

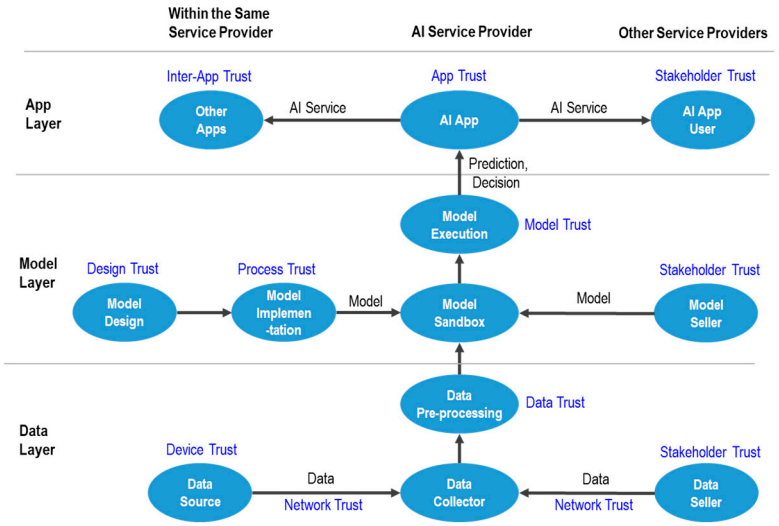

Figure 1. Trust managed objects in the AI hierarchy.

4.1.1. Data Layer

Operators that provide services based on AI systems install, operate, and manage data sources such as IoT sensors and web services, collect data directly or purchase necessary data from data sellers through data exchanges.

It is impossible to recover data through preprocessing using a Data Pre-Producer if there are many errors or omissions in the data acquired through collection and purchase. When low-quality data is provided to an AI model, it in turn leads to erroneous reasoning by the AI model, while intentionally biased data may be supplied, and in this case, the AI model is biasedly trained, which may also lead to erroneous results.

Therefore, quality management of collected and purchased data is a key prerequisite for the normal operation of the AI model, and analysis and management of trust targets in three aspects are required.

First, in the case of self-collection by installing a data source such as an IoT sensor, the data source should be regarded as the object of trust and trust properties and indicators should be defined and managed. For example, errors or missing data may be supplied due to the aging or hacking of IoT sensor nodes; therefore, it is necessary to continuously monitor and manage the quality of data generated by each data source.

Second, if third party data are purchased through data exchange, the third party data seller can be viewed by the purchaser as a major trust target. At this time, trust analysis and management of the data seller is required in terms of purchasing data since the third party's data collection and management is under the jurisdiction of the data seller.

Third, data quality degradation due to errors occurring in the process of transmitting data from the data source or seller to the data collector should be considered. In particular, there is a high probability of errors occurring in the process of wireless data transmission in ships and factories divided by steel walls, etc., and trust in the data transmission network must be analyzed and managed.

4.1.2. Model Layer

There are several trust targets and management elements even in the model layer, which can be said to be the core of an AI system. The AI system model can be supplied through self-development and purchase, and the trust target and management method for each case can be set differently. First, trust requirements must be reflected from the design stage of the AI model when developing an AI model on its own. In the "so-called" trust by design method, it should be tracked and managed whether key trust attributes in the model aspect, such as explainability and traceability, are reflected in the model design. In addition, verification of whether the trust-based model design is properly implemented should be performed in the implementation stage.

In the case of purchasing and applying a model supplied by an external developer rather than a self-developed model, the trust of the model seller who develops and supplies the model should be analyzed and managed.

In the case of AI models, erroneous results may be derived due to error data and biased data supplied for training, therefore continuous monitoring and verification of the developed and operating model should be accompanied.

4.1.3. Application Layer

The learned AI model can either be installed and operated on smart devices or applied to services. AI-based applications to which the AI model is applied collaborate with other applications within the same operator or provide intelligent services to general users, such as smartphone users. It is necessary to analyze and manage the trust between the AI application itself and the counterpart node when an AI-based application interacts with other applications or users.

4.2. Artificial Intelligence Systems Trust Framework

This section proposes a trust framework architecture for the AI system based on the trust requirements and characteristics of the AI system derived so far, trust analysis and management of AI systems are largely performed using Trust Agent, Trust Analysis, and Trust Information Management functional blocks as shown in Figure 2.

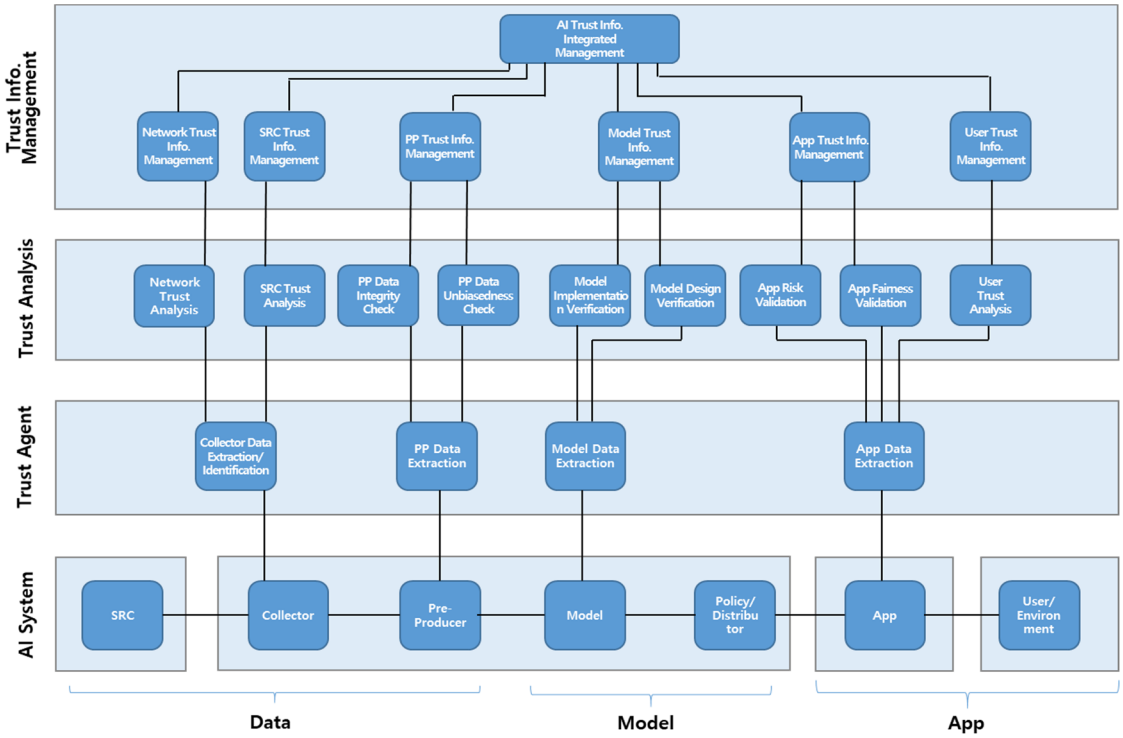

Figure 2. Trust framework for artificial intelligence systems.

The Trust Agent function block extracts and classifies data and logs required for trust analysis and management from C (Collector), PP (Pre-Producer), M (Model), and SINK functions of the AI system. Furthermore, it is responsible for transferring these trust-related data to the Trust Analysis block.

Based on trust-related data transmitted from the Trust Agent, the Trust Analysis function block includes functions to analyze and verify various aspects of trust attributes and indicators for SRC, PP, M, and SINK.

The Trust Information Management function block maintains and manages trust information analyzed through the Trust Analysis block. The trust index for each element of an AI system may change over time. The Trust Information Management function block needs to track and observe the history of how long the AI system elements have operated without problems in trust analysis and to continuously record and manage this information in the database. In addition, it takes the role of integrated management since the trust index for each element constituting the AI system is related to each other.

The configuration and operation of detailed trust verification and management functions for each element constituting the AI system follow in the next sections.

4.2.1. Trust Analysis and Management in Collector

Trust management is required to ensure that high-quality, error-free, and unbiased data is delivered from data sources and data vendors being collected as mentioned in Section 3.1. In addition, whether an error occurred in the transmission process through the network should be verified.

"Collector Data Extraction and Identification" extract the data required for trust analysis and verification for SRC and network through the interface with the Collector and delivers it to the "SRC Trust Analysis" and "Network Trust Analysis" functions.

"SRC Trust Analysis" analyzes the occurrence of errors and omissions in the data transmitted from each SRC, derives them as quantitative numbers and delivers them to the "SRC Trust Information Management" function. In this case, the SRC may be a data generating node such as an IoT sensor directly built by an AI system operator or a seller through data exchange.

The "Network Trust Analysis" function receives information related to retransmission and errors that occur in the data transmission and reception process from the network protocol functions connected to the Collector such as Transmission Control Protocol/Internet Protocol (TCP/IP), Constrained Application Protocol (COAP), and Message Queuing Telemetry Transport (MQTT). Furthermore, it is transmitted to the "Network Trust Information Management" function after analyzing the status information of the network channel. The "Network Trust Information Management" function manages to trust information about communication connection status and errors with the SRC over time.

4.2.2. Trust Analysis and Management in Data Pre-Producer (PP)

PP recovers errors and missing data through preprocessing of the collected data and provides it to the model. The "PP Data Extraction" function is responsible for extracting samples of the data input to the model that are preprocessed through the interface with the PP and delivering them to the "PP Data Unbiasness Checker" and "PP Integrity Checker" functions.

In terms of PP, the main requirement for trust is the removal of bias and securing integrity, which are analyzed and processed in the "PP Data Unbiasness Checker" and "PP Integrity Checker", respectively. The results analyzed by these functions are transmitted to the "PP Trust Information Management" function, where the trust index for processing the level of data purification required by the PP is continuously monitored and managed.

4.2.3. Trust Analysis and Management in Model (M)

The model should be trusted analyzed and verified during the design and implementation phases as described in Section 3.1.

First, it should be designed so that the operating state of the model can be traced and explained (explainability) according to the trust by design paradigm. "Model Design Verification" is a function that verifies the design and can be implemented in conjunction with the AI modeling platform. An authentication system may be introduced to verify and determine whether or not to trust in the operation of a model provided from an outsourced model developer.

"Model Implementation Verification" receives the operational status and log information of the implemented and trained model from the "Model Data Extraction" function and verifies whether the model is normally developed according to the design. Trust information about the model provided from the "Model Design Verification" and "Model Data Extraction" functions is transferred and managed by the "Model Trust Information Management" function.

4.2.4. Trust Analysis and Management in SINK (App)

The developed and trained models are delivered, loaded, and operated in cloud computing, edge computing, and user terminals through Policy (P) and Distributor (D) functions. It is necessary to analyze and manage, in terms of trust, whether the AI model itself, which is applied to applications and services, operates normally and whether it can be safely used.

In the "App Fairness Validation" and "App Risk Validation" functions, the AI model is analyzed and judged to establish whether it operates by the trust goals in terms of fairness and risk in the application (App) and service environment, and trust information is transferred to and managed by the "App Trust Information Management" function.

On the other hand, it is necessary to analyze the trust of the user and the usage environment as well as whether the app equipped with the AI model is operating normally based on the input information and feedback from the user who uses the AI app, and the "User Trust Analysis" function takes on these roles and handles them.

Trust information about users and usage environments is maintained and managed in the "User Trust Information Management" function.

The "AI Trust Information Integrated Management" function receives the trust index from each element of the AI system and performs a comprehensive analysis and management role. Furthermore, it is required to effectively provide the trust anomaly for each element to the AI system manager through visualization, etc.

4.3. Trust Analysis Model for Artificial Intelligence Systems

Figure 3 shows the relationship among trust attributes, trust indicates, and trust indexes for finding composite trust index as a quantified value in line with the three-layered concept (i.e., data layer, model layer, and app layer). The trust index for each layer or the entire system of the AI system should be derived and managed based on the trust target and management elements in the data layer, model layer, and application layer constituting the AI system.

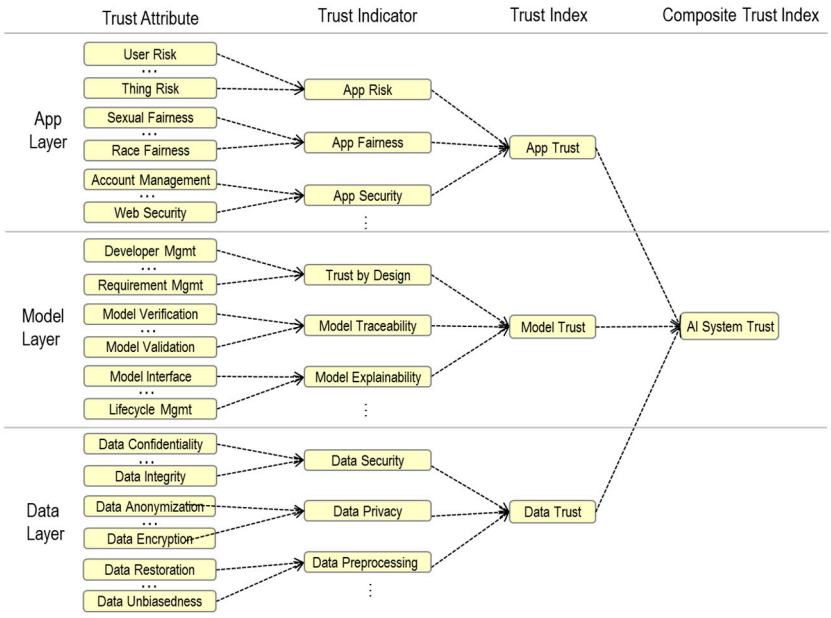

Figure 3. An example of an artificial intelligence trust analysis model.

To this end, the trust attribute and trust indicator for each layer should be defined as shown in Figure 3 based on the AI trust research trend analyzed in Section 2. A numerical trust indicator can be derived by applying ontology, statistics, and AI analysis techniques for trust properties (e.g., data unbiasedness and model validation.) based on AI model-related data and operation logs. A trust index for each layer is then calculated through weighting and aggregation. The trust index for the entire AI system can be derived by combining trust values for each layer, while for detailed methods, substantial additional research must be conducted according to the characteristics of the AI model and data.

5. Challenges for Standardization of AI Trust Management

By internationally standardizing the AI-specific trust management framework, global principles that can increase the stability and transparency of AI systems can be presented. Therefore, this section describes key items and technical issues required for AI trust management standardization.

5.1. In-Depth Understanding of AI Trust and Its Core Technologies

A clear definition of trust that can be interpreted differently depending on the situation must be a prerequisite to promote standardization of key technologies for AI trust, in the relationship with stakeholders, including developers and service users related to AI.

A more accurate understanding is needed in terms of broad trust in the overall ecosystem related to AI such as safety and accuracy, rather than the conventional narrow

aspects of security and privacy. Therefore, it is possible to design necessary element technologies or functional components only when the characteristics or properties of AI trust technology can be understood based on this.

5.2. Trust by Design Applied Trust-Based Lifecycle Operation Model Design

In applying AI technology, trust technology is required in the entire cycle, from the analysis process to obtaining the desired result through data collection and learning, and the actual action based on the obtained result to operate in the desired direction. A trustworthy AI operation model should be created from the initial design based on the "trust by design" concept.

5.3. AI Trust Reference Model

From a structural point of view, AI trust technology should be designed to support trust in all AI core functional blocks rather than being a technology that is limited to a specific layer. In this respect, to define detailed functions based on the reference model and to develop a standard that specifies related procedures, it is necessary to develop a universally applicable AI trust reference model through requirements analysis.

When we define the AI reference model that presents the linkage between the core functional blocks of the entire AI ecosystem, from the lower physical system to the upper control management to support AI, the main objective is to present a structure in which AI functions and trusts functions are interconnected.

5.4. An Artificial Intelligence Model That Evolves and Develops Transparently and Autonomously with Humans

It is important to autonomously evolve and fairly and clearly develop an AI model through interactions with people without being biased throughout the entire cycle, in order to support the trust of the model; this is the core technology of AI systems in an environment where people and technology coexist. To support this, a mechanism that can support transparency and fairness such as Explainable AI, which is a core attribute of trust and a trust support function, is required so that the model can evolve and develop autonomously in the right direction without being biased.

5.5. High-Reliability Application Support through Quality Control of Artificial Intelligence Models

There is a need for a standardized quality control technique that allows the AI model to determine whether data analysis can work safely and accurately derive optimal results. A quality management system for AI models that can be trusted to make decisions based on this should be established.

5.6. Artificial Intelligence Trust Analysis Mechanism

It is necessary to define a metric for trust and create a verifiable measurement technique to objectively derive trust analysis results. In addition, it is necessary to derive the values for the measurable parameters for trust mentioned above to determine the trust level in the physical and cyber environments and to define a computational trust analysis model that can be calculated considering the weight of the correlation between them. In particular, a standard for a composite trust index between data trust, model trust, and app trust that can define trust attributes and trust indicators according to AI data-model application and comprehensively evaluate reliability between them should be developed.

5.7. Risk Management System

A risk management model should be developed to identify and forecast the potential risks of AI and minimize them within the AI reliability framework. An overall risk management system should be established to continuously monitor risks and minimize the occurrence of problems by applying this risk management model, and take appropriate actions promptly in case of problems.

5.8. AI Trust Technology Verification and Certification

The various AI trust technologies presented so far must be verified and certified by an accredited institution so that the accredited trust technology can be spread. To this end, to be able to solve global interoperability issues, it is necessary to prepare standards for trust testing and test certification standards.

5.9. Artificial Intelligence Ethics and Social Issues

Various types of AI applications that can help people are being created through the interconnection between humans and devices. For this, a verification step for AI ethics is necessary. Furthermore, it is important to prepare an agreed standard for the trust support system that allows AI applications to be developed and utilized in the right direction after reviewing whether or not they can cause social problems such as various discriminatory issues.

6. Conclusions

After conducting a comprehensive analysis of various activities on trust and AI from a standardization perspective, this study derives trust management requirements for stable and predictable application and use of AI systems and their applications, as well as proposes a trust management framework based on them. The proposed trust management structure can be used to collect and analyze trust-related data throughout the entire lifecycle of the AI system, which consists of data collection, preprocessing, model, and application. This study also presents trust indicators and their properties required for AI trust analysis and proposes a trust model that can derive arithmetic measurable trust indices through their combination. Standardization of AI trust technology is a prerequisite for trust management technology to be applied to AI systems so that users can safely, reliably, and predictably use AI-based services. Therefore, in this study, strategies and challenges are presented to reflect the AI trust framework proposed through the analysis of existing standardization trends related to AI trust in the standardization. In addition to in-depth additional research on the AI trust model, future international standardization of AI trust based on this should be carried out.

Author Contributions: Conceptualization, T.-W.U. and G.M.L.; methodology, J.K. and G.M.L.; investigation, T.-W.U., J.K. and G.M.L.; writing—original draft preparation, T.-W.U. and G.M.L.; writing—review and editing, J.K., S.L. and G.M.L.; project administration, T.-W.U.; and funding acquisition, T.-W.U. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) (No. 2021-0-02068, Artificial Intelligence Innovation Hub, 50% and No.2022-0-01032, Development of Collective Collaboration Intelligence Framework for Internet of Autonomous Things, 50%).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Focus Group on AI for Autonomous and Assisted Driving (FG-AI4AD). Available online: https://www.itu.int/en/ITU-T/focusgroups/ai4ad/Pages/default.aspx (accessed on 1 April 2022).
2. Focus Group on Environmental Efficiency for Artificial Intelligence and other Emerging Technologies (FG-AI4EE). Available online: https://www.itu.int/en/ITU-T/focusgroups/ai4ee/Pages/default.aspx (accessed on 1 April 2022).
3. Focus Group on "Artificial Intelligence for Health". Available online: https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/default.aspx (accessed on 1 April 2022).
4. Lui, K.; Karmiol, J. *AI Infrastructure Reference Architecture*; IBM Systems: Armonk, NY, USA, 2018.
5. 6 Open-Source AI Frameworks You Should Know about. Available online: https://www.cmswire.com/digital-experience/6-open-source-ai-frameworks-you-should-know-about/ (accessed on 1 April 2022).
6. Zhang, C.; Li, W.; Luo, Y.; Hu, Y. AIT: An AI-Enabled Trust Management System for Vehicular Networks Using Blockchain Technology. *IEEE Internet Things J.* **2020**, *8*, 3157–3169. [CrossRef]

7. Zhang, T.; Qin, Y.; Li, Q. Trusted Artificial Intelligence: Technique Requirements and Best Practices. In Proceedings of the 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Shenyang, China, 20–22 October 2021; pp. 1458–1462. [CrossRef]
8. Ağca, M.A.; Faye, S.; Khadraoui, D. A Survey on Trusted Distributed Artificial Intelligence. *IEEE Access* **2022**, *10*, 55308–55337. [CrossRef]
9. Gasser, U.; Almeida, V.A.F. A Layered Model for AI Governance. *IEEE Internet Comput.* **2017**, *21*, 58–62. [CrossRef]
10. New AI Can Guess Whether You're Gay or Straight from a Photograph. *The Guardian*, 7 September 2019. Available online: https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph (accessed on 1 April 2022).
11. *Recommendation Y.3052, Overview of Trust Provisioning in ICT Infrastructures and Services*; ITU: Geneva, Switzerland, 2017.
12. Truong, N.B.; Lee, G.M.; Um, T.; Mackay, M. Trust Evaluation Mechanism for User Recruitment in Mobile Crowd-Sensing in the Internet of Things. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 2705–2719. [CrossRef]
13. Jayasinghe, U.; Lee, G.M.; Um, T.; Shi, Q. Machine Learning Based Trust Computational Model for IoT Services. *IEEE Trans. Sustain. Comput.* **2019**, *4*, 39–52. [CrossRef]
14. Oleshchuk, V. A trust-based security enforcement in disruption-tolerant networks. In Proceedings of the 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Bucharest, Romania, 21–23 September 2017; pp. 514–517. [CrossRef]
15. uTRUSTit, FP7 Project, ICT-2009.1.4—Trustworthy ICT. Available online: https://cordis.europa.eu/project/id/258360 (accessed on 1 April 2022).
16. TRUSTe Privacy Certification Standards. Available online: https://trustarc.com/consumer-info/privacy-certification-standards/ (accessed on 1 April 2022).
17. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
18. Peters, D.; Vold, K.; Robinson, D.; Calvo, R.A. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Trans. Technol. Soc.* **2020**, *1*, 34–47. [CrossRef]
19. *Technical Report, Trust Provisioning for Future ICT Infrastructures and Services*; ITU: Geneva, Switzerland, 2016.
20. Jayasinghe, U.; Truong, N.B.; Lee, G.M.; Um, T.-W. RpR: A Trust Computation Model for Social Internet of Things. In Proceedings of the 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), Toulouse, France, 18–21 July 2016; pp. 930–937.
21. Social Internet of Thing. Available online: http://www.social-iot.org./ (accessed on 1 April 2022).
22. Atzori, L.; Iera, A.; Morabito, G.; Nitti, M. The Social Internet of Things (SIoT)—When social networks meet the Internet of Things: Concept, architecture and network characterization. *Comput. Netw.* **2012**, *56*, 3594–3608. [CrossRef]
23. Huang, F. Building social trust: A human-capital approach. *J. Inst. Theor. Econ.* **2007**, *163*, 552–573. [CrossRef]
24. Zheng, Y.; Zhang, P.; Vasilakosd, A.V. A survey on trust management for Internet of Things. *J. Netw. Comput. Appl.* **2014**, *42*, 120–134.
25. Hummer, W.; Muthusamy, V.; Rausch, T.; Dube, P.; El Maghraoui, K.; Murthi, A.; Oum, P. ModelOps: Cloud-Based Lifecycle Management for Reliable and Trusted AI. In Proceedings of the 2019 IEEE International Conference on Cloud Engineering (IC2E), Prague, Czech Republic, 24–27 June 2019; pp. 113–120. [CrossRef]
26. Tao, C.; Gao, J.; Wang, T. Testing and Quality Validation for AI Software–Perspectives, Issues, and Practices. *IEEE Access* **2019**, *7*, 120164–120175. [CrossRef]
27. Srivastava, B.; Rossi, F. Rating AI systems for bias to promote trustable applications. *IBM J. Res. Dev.* **2019**, *63*, 5:1–5:9. [CrossRef]
28. Curzon, J.; Kosa, T.A.; Akalu, R.; El-Khatib, K. Privacy and Artificial Intelligence. *IEEE Trans. Artif. Intell.* **2021**, *2*, 96–108. [CrossRef]
29. Joshi, G.; Walambe, R.; Kotecha, K. A Review on Explainability in Multimodal Deep Neural Nets. *IEEE Access* **2021**, *9*, 59800–59821. [CrossRef]