Trustworthy Intelligent Systems: An Ontological Model

J. I. Olszewska
School of Computing and Engineering, University of the West of Scotland, U.K.

Keywords: Intelligent Systems, Software Engineering, Ontological Domain Analysis and Modeling, Knowledge Engineering, Knowledge Representation, Interoperability, Decision Support Systems, Dependability, Transparency, Accountability, Trustworthiness, Unbiased Machine Learning, Explainable Artificial Intelligence (XAI), Trustworthy Artificial Intelligence (TAI), Beneficial AI, Ethical AI, Society 5.0.

Abstract: Nowadays, there is an increased use of AI-based technologies in applications ranging from smart cities to smart manufacturing, from intelligent agents to autonomous vehicles. One of the main challenges posed by all these intelligent systems is their trustworthiness. Hence, in this work, we study the attributes underlying Trustworthy Artificial Intelligence (TAI), in order to develop an ontological model providing an operational definition of trustworthy intelligent systems (TIS). Our resulting Trustworthy Intelligent System Ontology (TISO) has been successfully applied in context of computer vision applications.

1 INTRODUCTION

The ongoing growth of AI-based technologies (Michael and Orescanin, 2022) in the new Society 5.0 (Fukuyama, 2018) has raised the question about the trustworthiness of these intelligent systems (Nair et al., 2021), (Black et al., 2022).

The trust concept has been studied in context of Information Systems back to late 1990s /early 2000s (Shapiro and Shachter, 2002). In particular, McKnight and Chervany (2000) proposed a well-established interdisciplinary model of trust types, relying on beliefs and intentions, and being based on characteristics such as competence, predictability, integrity, and benevolence.

More recently, with the development of AI systems whatever under water, on the ground, or in the air (Athavale et al., 2020), for applications encompassing smart cities, smart manufacturing, and smart agriculture (Coeckelbergh, 2019), the trust and trustworthiness concepts have been further investigated, leading to the field of Trustworthy Artificial Intelligence (TAI) (Jain et al., 2020). TAI is comprehended as a multi-dimensional concept which definition is yet to be fully set (Ashoori and Weisz, 2019). While many works defined the trustworthiness mainly in terms of predictability (Bauer, 2021), some recent works include also aspects such as ethics, lawfulness, robustness (Gillespie et al., 2020), dependability, beneficence, understandability (Chatila et al., 2021), non-maleficence, fairness, non-discrimination (Thiebes et al., 2021), privacy, transparency, explainability (Li et al., 2021), responsibility, controllability, accountability, as well as societal & environmental well-being (Liu et al., 2021).

Therefore, an ontological approach to capture the TAI domain and the ontological modeling to formalize the TIS concept look a promising way forward. Indeed, ontologies aid to formally specify knowledge of a domain and have been proven to be useful in Software Engineering (SE) and Artificial Intelligence (AI) domains alike (Olszewska and Allison, 2018), (Olszewska, 2020), (Bayat et al., 2016), (Fiorini et al., 2017), (Olszewska et al., 2017), (Olszewska et al., 2022).

Some ontologies have been produced to contribute to the Web of Trust by conceptualizing notions such as *trust*, *trustee*, *trustor* (Viljanen, 2005), *belief*, *trust in belief*, *trust in performance* (Huang and Fox, 2006), *performability*, *predictability*, *security* (Cho et al., 2016), as well as *institution-based trust* and *social trust* (i.e. inter-person trust) (Amaral et al., 2019). However, these older ontologies do not handle AI-based systems and their related challenges.

Besides, some ontologies have been built for specific applications such as recommendation systems (Porcel et al., 2015) or cyber-physical systems (Balduccini et al., 2018). Thence, their scope is limited intrinsically, while their implementations are not available.

Olszewska, J.
Trustworthy Intelligent Systems: An Ontological Model.
DOI: 10.5220/0011552700003335
In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 2: KEOD*, pages 207-214
ISBN: 978-989-758-614-9; ISSN: 2184-3228

For the TAI domain specifically, only two ontologies have been developed so far (Filip et al., 2021), (Manziuk et al., 2021), and both body of knowledge are based on the ISO/IEC TR 24028:2020 standard (ISO/IEC, 2020).

On one hand, Filip et al. (2021) proposes three main concepts, namely, the *governance* one, the *stakeholder* one, and the *technical* one. The governance-based trustworthiness is characterized by *transparency*, *explainability*, *accountability*, and *certification*. The stakeholder-based trustworthiness is characterized by *ethics*, *fairness*, and *privacy*. The technical-based trustworthiness is characterized by *reliability*, *robustness*, *verifiability*, *availability*, *resilience*, *quality*, and *bias*. Since the scope of this ontology is to coordinate the development of different standards and to check their consistency and/or overlap, this ontology is actually a taxonomy, without any axiomatization or implementation.

On the other hand, Manziuk et al. (2021) have elaborated concepts such as *trustworthiness*, *vulnerability*, *threat*, *challenge*, *high-level concern*, *stakeholder* and *mitigation measure*, where the latter concept has been decomposed in further sub-concepts such as *transparency*, *explainability*, *controllability*, *bias reduction*, *privacy*, *reliability*, *resilience*, *robustness*, *fault reduction*, *safety*, *testing*, *evaluation*, *use*, and *applicability*. Whereas this work presents an attempt to formalize the above-mentioned concepts of (ISO/IEC, 2020), it does not provide their implementation in any ontological language.

Therefore, as far as we are aware, none of the works which can be found in the current literature presents both conceptual and implementation models for TIS.

Hence, in this paper, we propose to capture the TAI domain in an OWL-based ontology we called Trustworthy Intelligent System Ontology (TISO) and to elaborate a formal and operational definition of the trustworthiness of intelligent systems (TIS).

In particular, to establish such formal and operational TIS definition, the TISO ontological concepts are encoded in Web Ontology Language Descriptive Logic (OWL DL), which is considered as the international standard for expressing ontologies and data on the Semantic Web (Guo et al., 2007), and uses Protege tool in conjunction with the FaCT++ reasoner (Tsarkov, 2014).

Besides, to be operational, a definition needs also to be measurable (Garbuk, 2018), (Cho et al., 2019). Thus, we worked on identifying measurable, attributes of trustworthy intelligent systems, while keeping a trade-off between software quality requirements (Nandakumar, 2022), (Mashkoor et al., 2022) and metric overabundance avoidance (DeFranco and Voas, 2022). As a result, we defined two initial, core TIS sub-concepts, namely, *dependability* and *transparency*.

It is worth noting that in context of specific applications or different perspectives such as user-centric, designer-centric, or regulator-centric, our TIS ontological modelling allows the core set of the TAI attributes to be expanded accordingly, e.g. to deal with the human-in-the-loop paradigm (Calzado et al., 2018) and related concepts (Kaur et al., 2023).

On the other hand, since the trustworthiness has also a temporal dimension as (Ashoori and Weisz, 2019), (Bauer, 2021), (Thiebes et al., 2021), (Kaur et al., 2023), our proposed TIS definition is further formalized using temporal logic in a way to represent and measure TIS over time and/or at different stages of the software development life cycle (Olszewska, 2019a).

The contributions of this paper is thus twofold. On one hand, we propose our TISO ontology which purpose is to aid in the trustworthiness assessment of AI-based systems. On the other hand, we introduce a fully operational definition of IV trustworthiness which is both formal and measurable and which has been implemented within TISO.

The paper is structured as follows. Section 2 presents the scope and the development of our Trustworthy Intelligent Systems' Ontology (TISO), while its evaluation and documentation are described in Section 3. Conclusions are drawn up in Section 4.

2 PROPOSED TISO ONTOLOGY

To develop the AI-T ontology, we followed the ontological development life cycle (Gomez-Perez et al., 2004) based on the Enterprise Ontology (EO) Methodology (Dietz and Mulder, 2020), since EO is well suited for software engineering applications (van Kervel et al., 2012), (Olszewska, 2019b).

The adopted ontological development methodology consists of four main phases, which cover the whole development cycle, as follows:

1. identifications of the scope of the ontology (Section 2.1);
2. ontology building which consists of three parts: the capture to identify the domain concepts and their relations; the coding to represent the ontology in a formal language; and the integration to share ontology knowledge (Section 2.2);
3. evaluation of the ontology to check that the developed ontology meets the scope of the project

(Section 3.1);

4. documentation of the ontology (Section 3.2).

2.1 Ontology Scope

The scope of this TISO ontology is, as follows:

- *TIS domain capture*: to identify the core concepts of the trustworthy artificial intelligence domain based on previous works and standards;
- *TIS guidelines*: to aid in building, testing, and deploying TIS;
- *TIS formalization*: to elaborate an operational (formal and measurable) definition of trustworthy intelligent systems;
- *TIS quantification*: to check if an intelligent system is trustworthy.

2.2 Ontology Building

In this work, we have studied further the challenges of modern trustworthy AI (TAI) to propose an operational definition for trustworthy IS (TIS). Hence, from the literature on the TAI domain and on the TAI-related ontologies (see Section 1), we have identified two core attributes of TAI systems in order to base TIS operational definition on.

On one hand, we have selected the *dependability* attribute. For that, we adopt the well-established definition proposed by (Avizienis et al., 2004) which considers dependability in terms of *reliability*, *maintainability*, *safety*, and *security*, with the security attribute encompassing itself the concepts of *availability*, *integrity*, and *confidentiality*.

On the other hand, we have chosen the *transparency* (Winfield et al., 2021) attribute. The transparency attribute, as defined in IEEE 7001:2021 (Winfield et al., 2021) for intelligent and autonomous systems, presents up to five levels of transparency, which definitions depend on the stakeholder status (e.g. user, designer, regulator, etc.). It is worth adding that the transparency attribute also includes notions of *explainability* and *fairness* and, in some degree, aspects of *familiarity* and *usability*.

Indeed, these two core attributes have been proved to be measurable according to the IEEE 982.1:2005 (IEEE, 2005) and IEEE 7001:2021 (Winfield et al., 2021) standards, respectively.

Moreover, TISO relies on further metrics which are specifically dedicated to AI systems. In particular, for the reliability attribute, in addition of the general reliability metrics which are described in IEEE 1633:2016 standard (Neufelder et al., 2015), TISO comprehends specific metrics to measure AI-based systems' reliability, as explained in (Olszewska, 2019b). Metrics to measure system safety are exposed in works such as (de Niz et al., 2018), and metrics to measure system security can be found e.g., in (Alanen et al., 2022). Besides, (Pressman, 2010) present extensively metrics for different attributes such as maintainability, while metrics for the transparency attribute have been studied, among others, by (Spagnuelo et al., 2016).

Next, we have coded the formal TIS knowledge in Descriptive Logic (DL). Thence, the concept of $Trustworthy\_Intelligent\_System$ is defined in DL, as follows:

$$\begin{aligned} Trustworthy\_Intelligent\_System \sqsubseteq\ & Intelligent\_System \\ & \sqcap \exists hasAttribute_{=Dependability} \\ & \sqcap \exists hasAttribute_{=Transparency}, \end{aligned} \tag{1}$$

where the $Dependability$ concept is defined in DL, as follows:

$$\begin{aligned} Dependability \sqsubseteq\ & Core\_Attribute \\ & \sqcap \exists hasAttribute_{=Maintainability} \\ & \sqcap \exists hasAttribute_{=Reliability} \\ & \sqcap \exists hasAttribute_{=Safety} \\ & \sqcap \exists hasAttribute_{=Security}, \end{aligned} \tag{2}$$

and the $Transparency$ concept is defined in DL, as follows:

$$\begin{aligned} Transparency \sqsubseteq\ & Core\_Attribute \\ & \sqcap \exists Stakeholder.H \\ & \sqcap \exists System.S \\ & \sqcap \exists Transparency\_Level_{S,H,i}, \end{aligned} \tag{3}$$

It is worth noting that the *System* concept can be defined in DL, as follows:

$$\begin{aligned} System \equiv\ & SubSystem.S_1 \\ & \sqcup \ldots \sqcup SubSystem.S_s, \end{aligned} \tag{4}$$

where S_j is the j^{th} sub-system of the system S, with $j \in \{1,...,s\}$, while the system's transparency level is defined as in IEEE 7001:2021 standard (Winfield et al., 2021). Therefore, we formalize the concept of $Transparency\_Level_{S,H,i}$ in DL, as follows:

$$\begin{aligned} Transparency\_Level_{S,H,i} \equiv\ & Transparency\_Level_{S,H} \\ & \sqcap \exists transparency\_level_{S,H}.value_{=i\in\{1,...,5\}}. \end{aligned} \tag{5}$$

If the inspection is carried out at a sub-system level S_j, the overall level of transparency of the system $Transparency\_Level_{S,H,i}$ could be rather defined in DL, as follows:

$$\begin{aligned} Transparency\_Level_{S,H,i} \equiv\ & Transparency\_Level_{S,H} \\ & \sqcap \exists transparency\_level_{S,H}.value_{=\min_i\{L_{S_j,H,i_j}\}}, \end{aligned} \tag{6}$$

with L_{S_j,H,i_j}, the level of transparency of the subsystem S_j, where $i_j \in \{1,...,5\}$ and where $j \in \{1,...,s\}$.

On the other hand, TIS properties such as *isTrustworthy* could be formalized in DL, as follows:

$$\begin{aligned} isTrustworthy \sqsubseteq\ & Intelligent\_System\_Property \\ & \sqcap \exists isDependable \\ & \sqcap \exists isTransparent, \end{aligned} \tag{7}$$

where *isDependable* property could be formalized in DL, as follows:

$$\begin{aligned} isDependable \sqsubseteq\ & System\_Property \\ & \sqcap \exists isMaintainable \\ & \sqcap \exists isReliable \\ & \sqcap \exists isSafe \\ & \sqcap \exists isSecure. \end{aligned} \tag{8}$$

Moreover, the *isReliable* property could be further formalized in DL, as follows:

$$\begin{aligned} isReliable \sqsubseteq\ & System\_Property \\ & \sqcap \exists System \\ & \sqcap \exists Reliability\_Metric \\ & \sqcap \exists hasReliabilityMetricValue_{=\{M_{R,S}.value \geq \theta_R\}}, \end{aligned} \tag{9}$$

with S, the system as defined in Eq. (4); R, the reliability attribute; $M_{R,S}$, a reliability metric; and θ_R, the threshold of this reliability metric.

In the Eq. (7), the *isTransparent* property could be further formalized in DL, as follows:

$$\begin{aligned} isTransparent \sqsubseteq\ & System\_Property \\ & \sqcap \exists hasTransparencyLevel_{=\{L_{S,H,i} \geq 1\}}, \end{aligned} \tag{10}$$

where $L_{S,H,i}$ is defined by Eq. (5) or Eq. (6), depending of the level of inspection of the system S.

Besides, the system's trustworthiness over time could be formalised using temporal-interval logic relations as introduced in (Olszewska, 2016), as follows:

$$\begin{aligned} isTrustworthyOverTime \sqsubseteq\ & Intelligent\_System\_Property \\ & \sqcap Temporal\_Property \\ & \sqcap (\diamond t_k)(\diamond t_l) \\ & (t_{k^+} < t_{l^-}) \\ & (isTrustworthy(S_k@t_k)) \\ & (isTrustworthy(S_l@t_l)) \\ & (M_{A,S_l}.value@t_l \geq M_{A,S_k}.value@t_k) \\ & \cdot (S_k@t_k \sqcap S_l@t_l), \end{aligned} \tag{11}$$

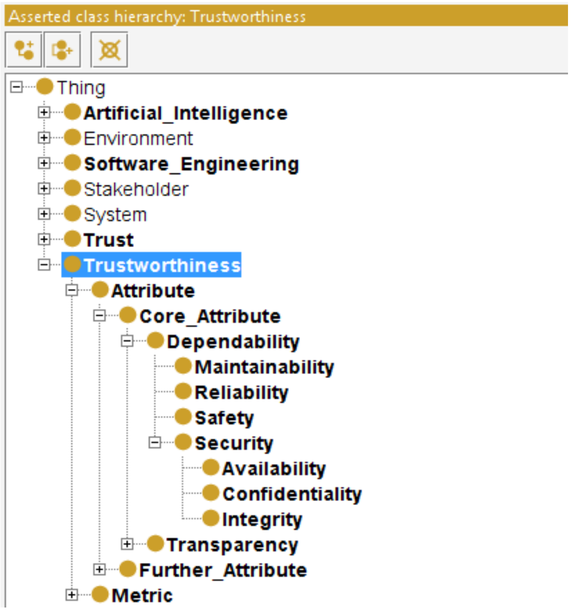

Figure 1: Main classes of the TISO ontology.

where M_{A,S_k} and M_{A,S_l} are the metrics of the attribute A in the time interval t_k and t_l, respectively; and where the attribute $A \in \{X,R,F,C,N\}$, with X, the *Transparency* attribute, R, the *Reliability* attribute, F, the *Safety* attribute, C, the *Security* attribute, and N, the *Maintainability* attribute. Furthermore, in Eq. (11), the temporal DL symbol $\diamond$ represents the temporal existential qualifier, and a time interval is an ordered set of points $T = \{t\}$ defined by end-points t^- and t^+, such as $(t^-,t^+) : (\forall t \in T)(t > t^-) \wedge (t < t^+)$.

These TISO ontological concepts and relationships have then been implemented in the Web Ontology Language (OWL) language, which is the language of all the software testing ontologies (Ferreira de Souza et al., 2013), and uses Protege v4.0.2 Integrated Development Environment (IDE) with the inbuilt FaCT++ v1.3.0 reasoner (Tsarkov, 2014) to check the internal consistency and to perform automated reasoning on the terms and axioms. An excerpt of the encoded concepts is presented in Fig. 1.

3 VALIDATION AND DISCUSSION

The developed TISO ontology has been evaluated both quantitatively and qualitatively in a series of experiments as described in Sections 3.1, while its documentation is mentioned in 3.2.

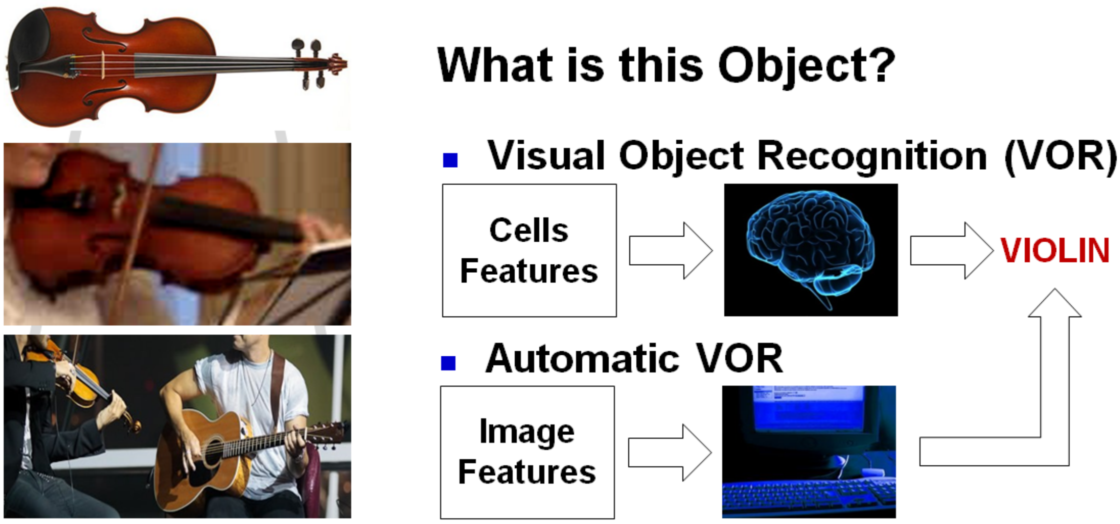

Figure 2: Overview of the visual object recognition (VOR) process.

3.1 Ontology Evaluation

To assess the TISO ontology on a real-world case study, we have analysed a computer-vision system (Olszewska, 2019b) which is an intelligent system processing visual inputs such as images or videos in order to extract high-level information that can be of interest for intelligent agents or humans alike. In particular, we have studied a computer-vision system dedicated to the visual object recognition (VOR).

The VOR process is illustrated on Fig. 2. (right side) and consists in processing the information from image features thanks to artificial intelligence methods in order to find the semantic label corresponding to the object of interest (i.e. the object to recognize/ the recognized object). The automated VOR process is similar to the natural VOR process, but presents a range of challenges, as illustrated on Fig. 2 (left side).

Hence, at first, on the picture on the top left corner of Fig. 2, one can see clearly the object of interest (in this case, a violin), because this object of interest is not rotated/has no deformations, and its foreground is not occluded, while its background is not noisy; the picture being centered on that object of interest – the object of interest being the only object visible in the scene. However, on the picture in the middle left side of Fig. 2, one can see that the object of interest (i.e. the violin) is rotated compared to the first picture. Moreover, in the second picture, the foreground is occluded, and the background is cluttered; the picture's overall resolution being poor as well – all these presenting additional challenges to the automated VOR process. Furthermore, on the picture on the bottom left corner of Fig. 2, one can see an even more complex background, since there are several objects such as one violin on the left, one guitar on the right and other objects around. The object of interest (i.e. the violin) is even more occluded, and it is not centered (but it is on the left side of the scene).

The object of interest has also a very different orientation and a different colour compared to the first and second pictures, leading to 'intra-class dissimilarities'. On the other hand, the colour of the violin on the left is similar to the colour of the guitar on the right of the picture, leading to 'inter-class similarities'. Please note that the word 'class' is used here as in the AI domain, rather than an OWL 'class'.

Thus, a computer-vision system handling the automated VOR needs to address the above-mentioned challenges. To study the trustworthiness of such intelligent vision system (IVS), experiments have been carried out using Protege v4.0.2 IDE and applying FaCT++ v1.3.0 reasoner on the TISO ontology together with the STVO ontology (Olszewska and McCluskey, 2011). It is worth noting that STVO stands for Spatio-Temporal Visual Ontology and comprises concepts such as object colour, object shape, object size, etc.

In particular, we have evaluated the different metrics M_{A,S_t} for the different attributes of the trustworthy intelligent systems (TIS) at different phases of the IVS software development life cycle (SDLC) (Olszewska, 2019a). In this way, we can apply the formal and operational definitions of the TIS concepts and properties as explained in Section 2 and their temporal dimension through these different stages of the life cycle of a VOR-dedicated IVS.

To exemplify this type of experiments, we can consider e.g. the Reliability attribute R and a reliability metric $M_{R,S_{D_n}}$ such as the VOR accuracy at four stages such as Describe (D3), Develop (D5), Deploy (D7), and Deploy' (D7') of the IVS SDLC, where D7 is the IVS deployment phase during the time interval t_{D7}, while D7' is the phase where the IVS system has been deployed for a time interval $t_{D7'}$ with $t_{D7^+} > t_{D7'^+}$. Let us assume the VOR accuracy threshold $\theta_R \geq 90\%$, so TISO helps to check if at all these Dn stages, $M_{R,S_{D_n}}.value \geq \theta_R$ as per Eq. (9) and $M_{R,S_{D_{n+1}}}.value \geq M_{R,S_{D_n}}.value$ as per Eq. (11). Once repeated for all the identified attributes and their related metrics and thresholds, this leads to contribute to the verification if the system 'isTrustworthy' and if the system 'isTrustworthyOverTime', respectively.

It is worth noting that measuring the trustworthiness at different SDLC stages generates different types of information about the trustworthy intelligent system. For example, at the D3 stage, the values of $M_{A,S_{D3}}$ can be incorporated in the system's requirements, leading to trustworthiness by design (Hamon et al., 2022). At the D5 stage, the values of $M_{A,S_{D5}}$ are obtained by running the system to test it and can be part of a certification process (Fisher et al., 2021). At the D7 stage, the values of $M_{A,S_{D7}}$ are reflecting the system performance as well as trustworthiness in real-world environment, while at the D7' stage, the values of $M_{A,S_{D7'}}$ indicate if an intelligent system is trustworthy over time, in a model-agnostic way (i.e. independently of the used machine learning/deep learning/.../logic techniques). Indeed, these measurements are performed on an intelligent system under the assumption that the intelligent system is a 'black-box' or a set of 'black-boxes', depending of the level of inspection. Therefore, our proposed model allows to assess the trustworthiness of intelligent systems, whatever AI approach (i.e. symbolism, connectivism, etc.) they use in their core process.

3.2 Ontology Documentation

The TISO ontology has been documented in Section 2. It is a middle-out, domain ontology that has been developed for trustworthy intelligent systems using EO methodology and that is not dependent of any particular software/system/agent/service/application/project. While TISO is based on the software engineering body of knowledge as well as the TAI principles and metrics, it has essentially been built on non-ontological resources such as primary sources (e.g. IEEE standards) – TISO having not reuse any existing TAI ontology.

On the other hand, the TISO ontology could be used in conjunction with other ontologies such as the core ontology for autonomous systems (CORA) (IEEE, 2015) or other robotics and automation ontologies (Fiorini et al., 2017) for further integration (Olszewska et al., 2017) within robotic and/or autonomous systems.

Moreover, the TISO ontology can be integrated with ontologies such as AI-T (Olszewska, 2020) for the testing (Black et al., 2022), (Araujo et al., 2022) as well as the verification (Dennis et al., 2016), (Araiza-Illan et al., 2022). of intelligent systems.

Besides, TISO can be applied together with other IEEE standards such as IEEE 7010:2020 (Schiff et al., 2020) or IEEE 7007:2021 (Prestes et al., 2021) to support Beneficial AI and/or Ethical/Legal AI, respectively.

4 CONCLUSIONS

In this paper, we have presented (a) a new ontology called TISO capturing the Trustworthy Artificial Intelligence (TAI) domain and identifying the core attributes of the trustworthiness in context of AI; (b) a formal and operational definition of Trustworthy Intelligent Systems (TIS), allowing interoperability, modularity, and multi-dimensionality (by including the temporal aspect). The developed ontology consists in integrating (b) within (a), in order to guide the development and use of trustworthy intelligent systems and to quantify the trustworthiness of intelligent systems over time and from various stakeholders' perspectives.

REFERENCES

Alanen, J., Linnosmaa, J., Malm, T., Papakonstantinou, N., Ahonen, T., Heikkila, E., and Tiusanen, R. (2022). Hybrid ontology for safety, security, and dependability risk assessments and security threat analysis (STA) method for Industrial Control Systems. *Reliability Engineering and System Safety*, 220:1–20.

Amaral, G., Sales, T. P., Guizzardi, G., and Porello, D. (2019). Towards a reference ontology of trust. In *Proceedings of OTM Confederated International Conferences On the Move to Meaningful Internet Systems*, pages 3–21.

Araiza-Illan, D., Fisher, M., Leahy, K., Olszewska, J. I., and Redfield, S. (2022). Verification of autonomous systems. *IEEE Robotics and Automation Magazine*, 29(1):2–3.

Araujo, H., Mousavi, M. R., and Varshosaz, M. (2022). Testing, validation, and verification of robotic and autonomous systems: A systematic review. *ACM Transactions on Software Engineering and Methodology*, pages 1–60.

Ashoori, M. and Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes.

Athavale, J., Baldovin, A., Graefe, R., Paulitsch, M., and Rosales, R. (2020). AI and reliability trends in safety-critical autonomous systems on ground and air. In *Proceedings of the Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops*, pages 74–77.

Avizienis, A., Laprie, J.-C., Randell, B., and Landwehr, C. (2004). Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing*, 1(1):11–33.

Balduccini, M., Griffor, E., Huth, M., Vishik, C., Burns, M., and Wollman, D. (2018). Ontology-based reasoning about the trustworthiness of cyber-physical systems.

Bauer, P. C. (2021). Clearing the jungle: Conceptualising trust and trustworthiness. In *Trust Matters: Cross-disciplinary Essays*, pages 1–16.

Bayat, B., Bermejo-Alonso, J., Carbonera, J. L., Facchinetti, T., Fiorini, S. R., Goncalves, P., Jorge, V., Habib, M., Khamis, A., Melo, K., Nguyen, B., Olszewska, J. I., Paull, L., Prestes, E., Ragavan, S. V., Saeedi, S., Sanz, R., Seto, M., Spencer, B., Trentini, M., Vosughi, A., and Li, H. (2016). Requirements for building an ontology for autonomous robots. *Industrial Robot*, 43(5):469–480.

Black, R., Davenport, J. H., Olszewska, J. I., Roessler, J., Smith, A. L., and Wright, J. (2022). *Artificial Intelligence and Software Testing: Building systems you can trust*. BCS Press.

Calzado, J., Lindsay, A., Chen, C., Samuels, G., and Olszewska, J. I. (2018). SAMI: Interactive, multi-sense robot architecture. In *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*, pages 317–322.

Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., and Yeung, K. (2021). Trustworthy AI. In *Reflections on artificial intelligence for Humanity*, pages 13–39.

Cho, J.-H., Hurley, P. M., and Xu, S. (2016). Metrics and measurement of trustworthy systems. In *Proceedings of the IEEE Military Communications Conference (MILCOM)*, pages 1237–1242.

Cho, J.-H., Xu, S., Hurley, P. M., Mackay, M., Benjamin, T., and Beaumont, M. (2019). STRAM: Measuring the trustworthiness of computer-based systems. *ACM Computing Surveys*, 51(6):1–47.

Coeckelbergh, M. (2019). Artificial intelligence: Some ethical issues and regulatory challenges. *Technology and Regulation*, pages 31–34.

de Niz, D., Andersson, B., and Moreno, G. (2018). Safety enforcement for the verification of autonomous systems. In *Proceedings of the SPIE International Conference on Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, page 1064303.

DeFranco, J. F. and Voas, J. (2022). Revisiting software metrology. *IEEE Computer*, 55(6):12–14.

Dennis, L. A., Fisher, M., Lincoln, N. K., Lisitsa, A., and Veres, S. M. (2016). Practical verification of decision-making in agent-based autonomous systems. *Automated Software Engineering*, 23(3):305–359.

Dietz, J. and Mulder, H. (2020). *Enterprise Ontology*. Springer.

Ferreira de Souza, E., de Almeida Falbo, R., and Vijaykumar, N. L. (2013). Ontologies in software testing: A systematic literature review. In *Proceedings of the Seminar on Ontology Research in Brazil*, pages 71–82.

Filip, D., Dave, L., and Harshvardhan, P. (2021). An Ontology for Standardising Trustworthy AI.

Fiorini, S. R., Bermejo-Alonso, J., Goncalves, P., Pignaton de Freitas, E., Olivares Alarcos, A., Olszewska, J. I., Prestes, E., Schlenoff, C., Ragavan, S. V., Redfield, S., Spencer, B., and Li, H. (2017). A suite of ontologies for robotics and automation. *IEEE Robotics and Automation Magazine*, 24(1):8–11.

Fisher, M., Mascardi, V., Rozier, K. Y., Schlinloff, B.-H., Winikoff, M., and Yorke-Smith, N. (2021). Revisiting software metrology. *Autonomous Agents and Multi-Agent Systems*, 35:1–65.

Fukuyama, M. (2018). Society 5.0: Aiming for a new human-centered society. *Japan Spotlight*, 27(5):47–50.

Garbuk, S. V. (2018). Intellimetry as a way to ensure AI trustworthiness. In *Proceedings of the IEEE International Conference on Artificial Intelligence Applications and Innovations*, pages 27–30.

Gillespie, N., Curtis, C., Bianchi, R., Akbari, A., and Fentener van Vlissingen, R. (2020). Achieving Trustworthy AI: A Model for Trustworthy Artificial Intelligence.

Gomez-Perez, A., Fernandez-Lopez, M., and Corcho, O. (2004). *Ontological Engineering*. Springer-Verlag.

Guo, Y., Qasem, A., Pan, Z., and Heflin, J. (2007). A requirement driven framework for benchmarking semantic web knowledge base systems. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):297–309.

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., and De Hert, P. (2022). Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, 17(1):72–85.

Huang, J. and Fox, M. S. (2006). An ontology of trust: Formal semantics and transitivity. In *Proceedings of the International Conference on lectronic commerce: The new e-commerce: innovations for conquering current barriers, obstacles and limitations to conducting successful business on the internet*, pages 259–270.

IEEE (2005). IEEE Standard Dictionary of Measures of the Software Aspects of Dependability: IEEE 982.1-2005.

IEEE (2015). IEEE Standard Ontologies for Robotics and Automation: IEEE 1872-2015.

ISO/IEC (2020). Information Technology - Artificial Intelligence - Overview of Trustworthiness in Artificial Intelligence: ISO/IEC TR 24028:2020.

Jain, S., Luthra, M., Sharma, S., and Fatima, M. (2020). Trustworthiness of artificial intelligence. In *Proceedings of the IEEE International Conference on Advanced Computing and Communication Systems*, pages 907–912.

Kaur, D., Uslu, S., Rittichier, K. J., and Durresi, A. (2023). Trustworthy Artificial Intelligence: A review. *ACM Computing Surveys*, 55(2):1–38.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., and Zhou, B. (2021). Trustworthy AI: From principles to practices.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A. K., and Tang, J. (2021). Trustworthy AI: A computational perspective.

Manziuk, E., Barmak, O., Krak, I., Mazurets, O., and Skrypnyk, T. (2021). Formal model of trustworthy artificial intelligence based on standardization. In *Proceedings of the International Workshop on Intelligence Information Technologies and Systems of Information Security*, pages 190–197.

Mashkoor, A., Menzies, T., Egyed, A., and Ramler, R. (2022). Artificial intelligence and software engineering: Are we ready? *IEEE Computer*, 55(3):24–28.

Michael, J. B. and Orescanin, M. (2022). Developing and deploying artificial intelligence systems. *IEEE Computer*, 55(6):15–17.

Nair, M. M., Tyagi, A. K., and Sreenath, N. (2021). The future with Industry 4.0 at the core of Society 5.0: Open issues, future opportunities and challenges. In *Proceedings of the IEEE International Conference on Computer Communication and Informatics*, pages 1–7.

Nandakumar, R. (2022). Quantitative quality score for software. In *Proceedings of the ACM Innovations in Software Engineering Conference*, pages 1–5.

Neufelder, A. M., Fiondella, L., Gullo, L. J., and Daughtrey, H. (2015). Advantages of IEEE P1633 for practicing software reliability. In *Proceedings of the IEEE Annual Reliability and Maintainability Symposium*, pages 1–6.

Olszewska, J. I. (2016). Temporal interval modeling for UML activity diagrams. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD)*, pages 199–203.

Olszewska, J. I. (2019a). D7-R4: Software development life-cycle for intelligent vision systems. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD)*, pages 435–441.

Olszewska, J. I. (2019b). Designing transparent and autonomous intelligent vision systems. In *Proceedings of the International Conference on Agents and Artificial Intelligence*, pages 850–856.

Olszewska, J. I. (2020). AI-T: Software testing ontology for AI-based systems. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD)*, pages 291–298.

Olszewska, J. I. and Allison, I. K. (2018). ODYSSEY: Software development life cycle ontology. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD)*, pages 303–311.

Olszewska, J. I., Barreto, M., Bermejo-Alonso, J., Carbonera, J., Chibani, A., Fiorini, S., Goncalves, P., Habib, M., Khamis, A., Olivares, A., Pignaton de Freitas, E., Prestes, E., Ragavan, S. V., Redfield, S., Sanz, R., Spencer, B., and Li, H. (2017). Ontology for autonomous robotics. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 189–194.

Olszewska, J. I., Bermejo-Alonso, J., and Sanz, R. (2022). Special issue on ontologies and standards for intelligent systems. *Knowledge Engineering Review*, 37:1–4.

Olszewska, J. I. and McCluskey, T. L. (2011). Ontology-coupled active contours for dynamic video scene understanding. In *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*, pages 369–374.

Porcel, C., Martinez-Cruz, C., Bernabe-Moreno, J., Tejeda-Lorente, A., and Herrera-Viedma, E. (2015). Integrating ontologies and fuzzy logic to represent user-trustworthiness in recommender systems. *Procedia Computer Science*, 55:603–612.

Pressman, R. S. (2010). Product metrics. In *Software engineering: A practitioner's approach*, pages 613–643. McGraw-Hill, 7th edition.

Prestes, E., Houghtaling, M., Goncalves, P. J. S., Fabiano, N., Ulgen, O., Fiorini, S. R., Murahwi, Z., Olszewska, J. I., and Haidegger, T. (2021). IEEE P7007: The first global ontological standard for ethically driven robotics and automation systems. *IEEE Robotics and Automation Magazine*, 28(4):120–124.

Schiff, D., Ayesh, A., Musikanski, L., and Havens, J. C. (2020). IEEE 7010: A new standard for assessing the well-being implications of artificial intelligence. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2746–2753.

Shapiro, D. and Shachter, R. (2002). User-agent value alignment. In *Proceedings of the AAAI International Conference on Artificial Intelligence*, pages 1–8.

Spagnuelo, D., Bartolini, C., and Lenzini, G. (2016). Metrics for transparency. In *Data privacy management and security assurance*, pages 3–18. Springer.

Thiebes, S., Lins, S., and Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, 31:447–464.

Tsarkov, D. (2014). Incremental and persistent reasoning in FaCT++. In *Proceedings of the OWL Reasoner Evaluation Workshop (ORE)*, pages 16–22.

van Kervel, S., Dietz, J., Hintzen, J., van Meeuwen, T., and Zijlstra, B. (2012). Enterprise ontology driven software engineering. In *Proceedings of the International Conference on Software Technologies (ICSOFT)*, pages 205–210.

Viljanen, L. (2005). Towards an ontology of trust. In *Proceedings of the International conference on trust, privacy and security in digital business*, pages 175–184.

Winfield, A. F. T., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., Muttram, R. I., Olszewska, J. I., Rajabiyazdi, F., Theodorou, A., Underwood, M. A., Wortham, R. H., and Watson, E. (2021). IEEE P7001: A proposed standard on transparency. *Frontiers in Robotics and AI*, 8:1–11.