Study on Test Plan Establishment for Enhancing Trustworthiness of Artificial Intelligence Systems

Yejin Shin*
AI Trustworthiness Verification Team
Telecommunications Technology Association(TTA)
Seongnam, South Korea
yepp1252@tta.or.kr

Sangyeon Kang
AI Trustworthiness Verification Team
Telecommunications Technology Association(TTA)
Seongnam, South Korea
cellina7702@tta.or.kr

***Abstract*—Compared to traditional software, artificial intelligence systems (AIS) include uncertainty in the inference results during the training process, which is one of the factors that could undermine the trustworthiness of the system. To overcome this uncertainty of AIS, a trustworthiness test plan establishment that accounts for the complexity and operational environment is required. In this study, four requirements that must be considered by the system designers, developers, and quality managers are proposed for establishing the test plan of the trustworthiness of AIS. If AISs satisfy the proposed requirements, it will help secure safety and transparency among the AI trustworthiness factors.**

***Index Terms*—artificial intelligence, AI, trustworthiness, trustworthy AI**

I. Introduction

Artificial intelligence (AI) technology is becoming common in daily life as it finds applications in various fields, such as autonomous driving and contactless chatbot services in banking. Artificial intelligence systems (AISs) learn from large-scale datasets and derive probabilistic results based on these data; thus, they can suffer from trustworthiness issues [1]. AIS, compared to traditional software, includes uncertainty for the inference results during the training process. This uncertainty is even more critical in applications where the AIS movement could harm humans or the environment [2], and it is one factor that could undermine the trustworthiness of the system. Therefore, in order to secure the trustworthiness of AIS, tests for confirming the trustworthiness of the AIS are also required in addition to tests for confirming the software quality.

Planning that considers the complexity and operational environment of the system is required to test the trustworthiness of AIS. Additionally, periodic and continuous tests must be performed throughout all lifecycle phases of AIS according to the established plan. Therefore, in this study, we argue that the establishment of an AIS trustworthiness test plan should be performed first in the 'planning and design' phase among the AI lifecycle phases, and specific activities to be performed are included in the proposal.

Previous studies [3], [4], [5] have described the verification and test method for improving the trustworthiness of AIS. However, there is a lack of research on methods for establishing a test plan for practitioners during the 'planning and design' phase. In this study, four requirements that the system designers, developers, and quality managers must consider for establishing the trustworthiness test plan for AIS are proposed. The research method involves the analysis of ISO/IED TR 29119-11:2020 [6] and the application of the results to the AI trustworthiness factors. If the AIS meets the proposed requirements, we believe that it will support the safety and transparency of the AIS among the AI trustworthiness factors.

This paper is structured as follows. In Section 2, we provide background information on AI trustworthiness, which is essential for understanding the proposed requirements. In Section 3, we outline the requirements for test environment design that should be considered to establish an AIS trustworthiness test plan. In Section 4, we propose requirements for a consultation system that includes the formation of an expert consensus group and a user evaluation group. These requirements aim to ensure that the complexity and operational environment of the AIS is taken into account during the test plan establishment. Finally, we conclude the paper in Section 5 by summarising the proposed requirements.

II. Background of AI trustworthiness

In ISO/IEC TR 24028 [7], trustworthiness was defined as the 'ability to meet stakeholders' expectations in a verifiable way. In addition, in the Ethics Guidelines for Trustworthy AI by the European Commission [8], a Trustworthy AI (TAI) was described to 'complying with all applicable laws and regulations(lawful), ensuring adherence to ethical principles and values(ethical), and both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm(robust)'. Therefore, AI trustworthiness must be recognised as an essential factor for the preemptive identification of the risk and side-effects of AI and a comprehensive preparation and verification from the aspects of technology, society, ethics, and so on.

Currently, research on the implementation of trustworthiness on AIS is vastly performed. These studies focused on five factors, including fairness, safety, transparency, accountability, and privacy. Works on fairness [1], [9], [10], [11] are focused on the method of preventing the discriminative output of AI. In addition, safety is related to testing the robustness of the training dataset or AI model from attacks [12], [13], [14]. Transparency researches [15], [16], [17] are relevant to the

TABLE I
FOUR REQUIREMENTS FOR ESTABLISHING THE TEST PLAN OF AIS TRUSTWORTHINESS

#	Category	Requirement
1	Test Environment Design	Was the AIS operational environment considered during test environment determination?
2		For AIS that requires a virtual test environment, was simulator prepared?
3	Organising Consultation System	Was a consultation system constructed for determining the expected output of AIS?
4		Was a user evaluation group constructed for identifying the explainability and interpretability?

description of the predictive process of the algorithm for overcoming the black box characteristics of AI.

In this study, four requirements that must be considered by practitioners, such as system designers, developers, and quality managers, when establishing the test plan for the trustworthiness of the AIS are proposed. Table 1 lists the four requirements for establishing the test plan for the trustworthiness of AIS. If these requirements are satisfied, it will support the partial securement of the safety and transparency of the above-mentioned factors.

III. REQUIREMENTS RELEVANT TO TEST ENVIRONMENT DESIGN

While designing a test environment, the features of the AI must be considered. For the AIS, the virtual test or actual environment test must be considered based on its complexity or level of risk. In particular, the ethics guideline of AI by UNESCO [18] recommends various tests for AIS with potential risks to human rights as a part of an ethical impact assessment by the parties of interest before its release, and a test must be performed in an environment identical to the actual circumstance if necessary.

While it is appropriate to perform a real environment test for accurate testing, the test must be performed within a reasonable time and cost range, and a real environment test may not be suitable for a system with highly complex operational conditions. In addition, applying a real environment test to an AI that physically interacts with humans could lead to dangerous situations. Considering this, a virtual test must be carried out. Therefore, it is necessary to design a test environment after determining the appropriate test environment by considering the system characteristics. The examples of the considerations presented by the ISO/IEC TR 29119-11:2020 [6] standard and others for the test environment design are as follows:

- Does the operational environment of AIS change constantly and is it complex?
- Is it a system with potential risks to human rights?
- Is the test performable within a reasonable times and cost range?
- Are individuals in the environment during real environment testing (e.g., vehicles, buildings, animals, and humans) harmed?

A. Was the operational environment of AIS considered during test environment determination?

For AISs with numerous parameters, such as the operational environment restrictions, variety in function, and performance degradation factors, the number of test suites can be nearly infinite. In this case, the use of pairwise testing as a method of combination testing, which can reduce the number of test suites through the combination of parameters, must be considered [19]. In contrast, AISs that include scenarios that risk damage to the individual in the environment during testing or that have difficulty in generating scenarios for edge cases must consider virtual testing environments. Moreover, a virtual test can be adopted for the case of real environment testing owing to the difficulty in preparing a test environment (e.g., robots that explore nuclear power plant accident sites) [6].

To apply the requirements proposed in this study to an autonomous driving AI, a virtual test using a simulator and an actual test for all scenarios must be considered, and different test environments must be determined based on the driving scenario. Additionally, the test environment must differ according to the target object and the circumstances included in the driving scenario. This is because, during lane change upon falling object recognition, the scenario may be tested in actual environments; however, stop scenarios based on the recognition of pedestrians cannot be tested in real environments.

B. For AIS requiring virtual test environments, was a simulator prepared?

In some domains, there are simulators published as open-source, which can be utilised for virtual testing if they are suitable for the AIS to be developed. However, if the simulator is not suitable or if there are no reusable simulators, constructing a new simulator is necessary. In this case, it is important to previously review the scale of additional resources for the construction (e.g., labour, costs, and time) for planning. Thus, for AISs that require virtual test environments, a resource reviewing process must be considered during the 'planning and design' phase. This will prevent the occurrence of additional costs and delay due to the unpredicted lack of resources during the simulator construction process and enhance the efficiency and accuracy of the virtual test. Furthermore, it must be identified whether the simulator is representative of the operational environment. For example, a high level of image representativeness is required for the pedestrian avoidance test of autonomous driving vehicles.

IV. REQUIREMENTS RELEVANT TO CONSULTATION SYSTEM

During the AIS test design, a consultation system necessary for the design must be constructed. Most AIS have low reproducibility owing to their high complexity, and transparency is difficult to secure. Additionally, the system complexity becomes problematic to the test oracle that determines the expected output. Therefore, it is difficult to determine whether the test has succeeded or failed.

If the system requires a description of the inference results of the AIS, the assessment criteria for the explainability may differ based on the target user that identifies the system output. Additionally, the assessment criteria for interpretability, which scales the extent of the understanding of the operational method of the AI, is also dependent on the target user. In other words, the required level for explainability and interpretability differs for each system or target user. Therefore, a consultation system necessary for the establishment of the assessment criteria of the explainability and interpretability of the system output or the decision on the expected output must be constructed, and it is appropriate to design the test through the derivation of a consensus among participants.

A. Was a consultation system for determining the expected output of AIS constructed?

The test oracle problem refers to the problem that occurs when there is a difference between the AIS training results and the actual output. A council of internal and external experts in the corresponding domain must be formed to overcome this issue and increase the trustworthiness of AIS. Here, the council plays the role of determining the expected output of the system, and it must be recognised that time may be required for several experts to agree on the expected output [6].

When comprising the council, it is important to consider the variety and experience of the experts necessary for determining the expected output of the system. Additionally, all possible aspects must be considered during the process of determining the expected output. Council experts may present different expected outputs for identical input. Therefore, an approval criterion must be previously defined for achieving a consensus among the experts. For instance, there is a method of approving when two out of three experts agree on a certain expected output.

B. Was a user evaluation group constructed for identifying the explainability and interpretability?

For systems that require a description on the AIS output, it is necessary to test the extent of the sufficiency of the system's explainability and interpretability. The test criteria will mainly be on the extent of how easily the AIS target user understands the output and operational method. The depth of necessary description may differ based on the field of AI application, and the method of description may have to differ based on the target user. For example, the explainability of an AIS that retouches photographed images must differ from an AIS that is utilised in diagnosing and determining operability.

Therefore, a user evaluation group must be formed to determine the level of explainability, and this must be reflected while implementing the model and system. Considering this, the target user must be clearly defined during the 'planning and design' phase before constructing a user evaluation group.

To apply the requirements proposed in this research to autonomous driving artificial intelligence, misunderstanding or misinterpretations from different drivers may occur in the case of the AIS providing an autonomous driving system output through the human-machine interface. Therefore, an evaluation group comprised of drivers from various backgrounds must be formed to include the process of identifying interpretability. Furthermore, preparing the criteria for success and failure of the test is also necessary based on the assessment results of the user evaluation group. For instance, this could include the preparation of the quantitative criteria, such as determining the success when the average score is equal to or above a certain number, or the preparation of the determination criteria, such as using the truncated mean during average score calculation.

V. Conclusions

The service provided by AISs is becoming increasingly important in the daily lives of people, and the trustworthiness issue is becoming a greater issue accordingly. Therefore, the importance of establishing a test plan for improving the trustworthiness of AIS and the method are discussed in this study. Particularly, the requirements that the system designers, developers, and quality managers must refer to for establishing the AIS trustworthiness test plan were presented based on ISO/IEC TR 29119-11:2020.

The requirements relevant to the consultation system construction along with those for the test environment design must be considered to improve the trustworthiness of AIS. If these requirements are met, the safety and transparency of the AIS among the trustworthiness factors can be secured. Therefore, we anticipate that the content proposed in this study supports the safeguarding of the trustworthiness of AIS.

Additionally, it is necessary to identify whether traditional attributes that apply to the previous software system and the attributes that correspond to the AI effectively apply. Thus, verification procedures in terms of system performance, security, and quality must be performed in parallel as well as those for the presented requirements in this study.

Acknowledgment

This work was supported by the Korean Ministry of Science and ICT (MSIT) as Establishing the foundation of AI Trustworthiness (TTA).

References

[1] Y. Shin, K. Cho, J. Kwak, and J. Hwang, "Development of a Method for Ensuring Fairness of an Artificial Intelligence System in the Implementation Process," 13th International Conference on Information and Communication Technology Convergence (ICTC), pp. 2192–2194, 2022.

[2] M. Kläs, and A. M. Vollmer, "Uncertainty in Machine Learning Applications: A Practice-Driven Classification of Uncertainty," In Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, pp. 431–438, 2018.

[3] B. Shneiderman, "Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems," ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 10, no. 4, pp. 1–31, 2020.

[4] M. Brundage, et al. "Toward trustworthy AI development: mechanisms for supporting verifiable claims," arXiv preprint arXiv:2004.07213, 2020.

[5] T. Dreossi, D. J. Fremont, S. Ghosh, E. Kim, H. Ravanbakhsh, M. Vazquez-Chanlatte, and S. A. Seshia, "VerifAI: A Toolkit for the Formal Design and Analysis of Artificial Intelligence-Based Systems," In Computer Aided Verification(CAV): 31st International Conference, pp. 432–442, 2019.

[6] ISO/IEC TR 29119-11:2020, Guidelines on the testing of AI-based systems, 2020.
[7] ISO/IEC TR 24028:2020, Overview of trustworthiness in artificial intelligence, 2020.
[8] European Commission, Ethics Guidelines for Trustworthy AI, 2019.
[9] J. Zhang, Y. Shu, and H. Yu, "Fairness in Design: A Framework for Facilitating Ethical Artificial Intelligence Designs," International Journal of Crowd Science, vol. 7, no. 1, pp. 32–39, 2023.
[10] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Towards measuring fairness in AI: the casual conversations dataset," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 4, no. 3, pp. 324–332, 2021.
[11] D. Varona, and J. L. Suárez, "Discrimination, Bias, Fairness, and Trustworthy A," Applied Sciences, vol. 12, no. 12, 5826, 2022.
[12] T. Ouyang, Y. Isobe, V. S. Marco, J. Ogata, Y. Seo, and Y. Oiwa, "AI robustness analysis with consideration of corner cases," 2021 IEEE International Conference on Artificial Intelligence Testing (AITest), pp. 29–36, 2021.
[13] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang, "Adversarial robustness: From self-supervised pre-training to fine-tuning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 699–708, 2020.
[14] J. Meng, F. Zhu, Y. Ge, and P. Zhao, "Integrating safety constraints into adversarial training for robust deep reinforcement learning," Information Sciences, vol. 619, pp. 310–323, 2023.
[15] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE transactions on neural networks and learning systems, vol. 32, no. 11, pp. 4793–4813, 2020.
[16] M. Vössing, N. Kühl, M. Lind, and G. Satzger, "Designing transparency for effective human-AI collaboration," Information Systems Frontiers, vol. 24, no. 3, pp. 877–895, 2022.
[17] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum,"What do we want from Explainable Artificial Intelligence(XAI)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," Artificial Intelligence, vol. 296, p. 103473, 2021.
[18] UNESCO, Recommendation on the ethics of artificial intelligence, 2021.
[19] D. R. Kuhn, D. R. Wallace and A. M. Gallo, "Software fault interactions and implications for software testing," IEEE Transactions on Software Engineering, vol. 30, no, 6, pp. 418–421, 2004.