Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p27910-cd

The importance of Human Reliability for assessing the Risk of Artificial-Intelligence-Systems according to the EU AI-Act

Prof. Dr. habil. Oliver Straeter

Arbeits- und Organisationspsychologie, Kassel University, Germany. E-mail: straeter@uni-kassel.de

The EU AI Act establishes a risk-oriented approach to the use of artificial intelligence within the European system. Socially relevant application scenarios, distinguishing four risk levels. Applications, such as lending, career opportunities, and autonomous systems (vehicles, robotics), are assigned to risk level two. This risk level is characterized by the fact that the application of an AI system requires additional human monitoring or reviewing. A risk-oriented approach necessitates risk-oriented modeling, which, with regard to the role of humans, requires a human reliability analysis (HRA). HRA methods are well-established in the assessment of high-risk systems, particularly in nuclear safety and aviation safety. The VDI addresses these aspects in VDI Guideline 4006, Parts 1-3. The appropriate application of HRA methods is crucial for the reliable implementation of the AI Act. Various aspects must be considered, which are discussed in this paper. This results in important requirements for the methods used to certify AI systems under the European AI Act, as well as for processes involved in the practical implementation of AI systems. This article will examine this issue, present the evaluation scenario for AI systems with human assistance, and discuss the resulting safety measures required for any use of AI systems use-cases as mentioned above.

Keywords: AI-act, human reliability, goal-oriented behaviour.

1. Introduction

The methodology of artificial intelligence (AI), particularly large-language models (LLM), is not only a fascinating topic in research but also a powerful tool for analyzing complex, heterogeneous, and extensive issues in business settings. This makes LLM ideally suited to provide methodological support for many applications, such as autonomous driving, personnel selection, financial decisions, and managing security issues, as these also involve processing complex, heterogeneous, and extensive data.

However, AI systems inherently have two issues: First, they have an issue of integrity in that they vary in their outputs even if the input stays stable. Second, they are prone to so-called hallucinations, as they generate new responses based on input data or "invent" suggestions or solutions that sound convincing but may not be covered by the input data. Integrity issues and hallucinations are particularly critical for all security-related considerations regarding site selection. Furthermore, responses are based on reference data and classifications that are usually unknown or hidden from the end user.

In order to not let these issues go through to unsafe states for technical systems or societal developments, the EU's AI act defines four risk levels to classify AI Applications (Figure 1) according to EU-Act (2025).

The EU's AI law would only permit an AI system with a more direct impact on safety (at least level two, high risk) in combination with a person verifying the outputs or statements.

According to the German Standard for Human Reliability published as VDI 4006, the tasks of verifying the outputs, when dealing with such AI assistance systems, can be performed in two modes by a person:

- in a task-oriented manner, i.e., a person uses the AI assistant in the best intended way and critically reflects outputs and builds his own independent view on a situation
- in a goal-oriented way, i.e. a person uses the AI assistant to shortcut processes and uses the output of an AI system as given and just forwards the results without own reflection

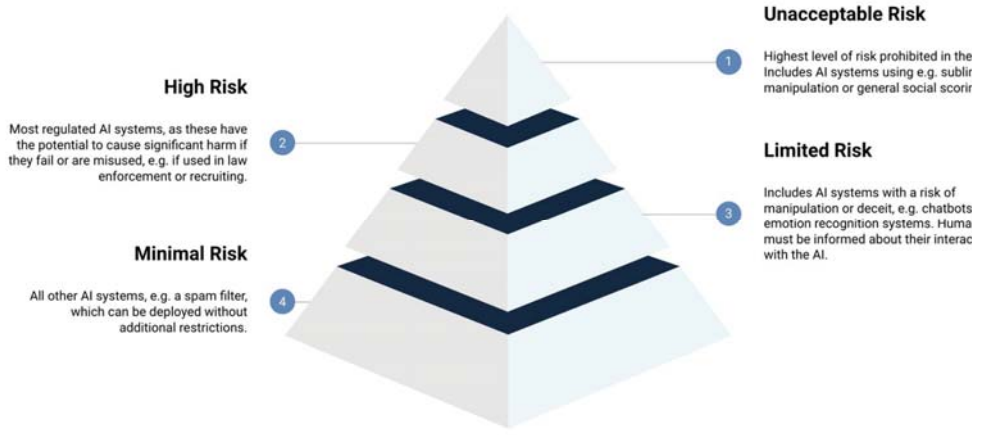

Figure 1: Risk-oriented assessment according to the EU Act on Artificial Intelligence (AI) / EU-Act (2025).

This distinction is based on considerations of Straeter (2019) on the state of the art of behavioral dimensions in Human Reliability Assessment.

From a human reliability standpoint, one cannot assume that the verification activity will be carried out in a task-oriented manner when such an AI system is used as an assistance. The securing function of a person is therefore likely not performed. Therefore, according to the AI law, AI systems would hence generally not be usable for the aforementioned use cases unless reliable human oversight is in place.

The following demonstrates how Human Reliability according to VDI 4006 assesses human reliability of an AI system on risk-level two. The outlook discusses open questions regarding human reliability and addresses possible future developments in the VDI 4006 standard. These requirements also play an important role in the field of security; therefore, this connection is methodically examined in the article.

2. Human Reliability Assessment Approach

The EU Act on Artificial Intelligence (AI) mandates a risk-based assessment of AI systems to ensure they do not have negative consequences for people or society. Developments in the United States show that product manufacturers may downplay risks in safety assessments to gain a stronger market position for their products. Most prominent example is the ‘autopilot’ of Tesla, which is forbidden to be called ‘autopilot’ in Germany because the autonomous system is not covering all functions of a real autopilot and misleads the driver in a mode of overconfidence of the autonomous car (Auto-Motor-Sport, 2024; Grabbe, 2024).

In general, any lack of fact-checking of the outputs of an AI system can lead to technically or socially harmful developments. Therefore, it is all the more important to conduct the risk assessment of an AI system according to the procedures commonly used for other industrial systems (Straeter, 2025).

The scenario to be assessed and judged upon is an AI system classified as an assistance system used by a person or group. This can only be in risk-level 2 if a Human reliable uses the system and not over-relies in the outputs without questioning whether the output is reasonable or not. If this function fails, the same AI system needs to be classified as risk-level one, hence having no operational allowance in the EU.

An example: An automated AI based personnel recruitment system would be classified according to the EU act as a risk-level one system and would not get any operational license. Only if the operator of the AI system ensures that a human being monitors, reflects and - if necessary - corrects the output, such system may get risk-level 2 and can be used.

Human reliability assessment plays a crucial role in this decision. What happens if human reliability assessment is not taking seriously, is demonstrated negatively by outnumbered examples including more recent major incidents such as the Boeing 737 Max crashes, accidents involving autonomous vehicles, human-robot collaboration, and highly automated process industries (nuclear power plants). However, the aspect of human reliability extends beyond the operation of technical systems to encompass their planning, design, construction, and decommissioning (e.g., Hollnagel, 1999 and 2012; Apostolakis et al., 2004).

The VDI (Association of German Engineers) addresses these aspects in VDI Guideline 4006, Parts 1-3. In addition, expert recommendations have been developed on specific aspects of human reliability: (1) interactions between humans and autonomous systems (including AI systems) and (2) managing human reliability in the operational context of occupational safety (VDI-EE-AS 2021, VDI-EE-VZ 2024). The significance of human reliability is thus broadly covered. The relevance of these guidelines in different technical domains is demonstrated in Straeter (2019).

In the following, the requirements of VDI 4006 are reflected upon in light of damage scenarios that can be expected with artificial intelligence systems. Faulty system states may arise because of preconceived notions (biases) prevail during the planning phase, leading to unfavourable system configurations or to inadequate use of AI systems not in accordance to the AI act requirement of monitoring, reflecting and questioning the outputs of such systems. It will be demonstrated that HRA assessments according to VDI 4006 (or similar EU guidelines on human reliability) are necessary in order to manage the risk of AI systems according to the EU AI act.

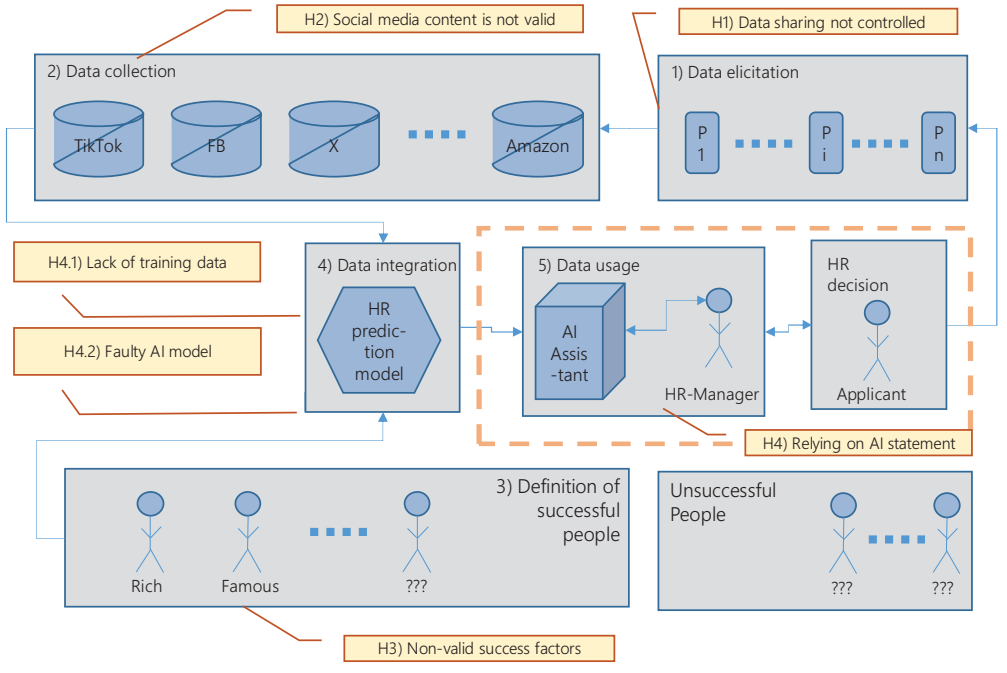

Figure 2: System structure of an AI system to support personnel selection (P for Person; H for Hazard).

3. Human Reliability Assessment of artificially intelligent systems

An AI system is an autonomous system. The reliability of an autonomous system's interaction with a human depends crucially on the extent to which scenarios exceeding its design parameters prevail and/or whether the human uses the assistance system either in a task-oriented or goal-oriented mode (Straeter, 2005 and 2019). Figure 2 shows an example of an AI system used to support personnel selection (dashed red box). A human resources manager uses AI assistance to make a hiring decision regarding an applicant. For the AI system to provide this assistance, a number of system functions are required: 1) data collection, 2) data storage, 3) definition of successful individuals, 4) data integration, and 5) data utilization. A number of hazards can be associated with these system functions. In Figure 2, these hazards are marked with H*).

The scope of consideration is crucial for evaluating an AI system. The essential function of the system is marked with the dashed red box. When evaluating the scenario according to VDI 4006 (2026) and VDI-EE-AS (2021) the fault tree as outlined in Figure 3 will be the result: AI systems designed as assistance systems may either be operated in a task-oriented or goal-oriented manner. In a task-oriented manner, the logic is that the system merely makes suggestions and leaves the final decision-making to the human. In terms of reliability, this ensures at least the level of human reliability in a task-oriented sense (i.e., on the order of E-03) and comes up to an overall reliability by independence of the AI system's reliability. However, in a goal-oriented sense, the human contribution to reliability would be virtually eliminated if the human blindly relied on the assistance (i.e., on the order of E-01).

The Reliabilities are calculated using the HRA-Method CAHR either for task- or goal-oriented behaviour (Straeter, 2000). The values represent the average expectation of reliabilities which might be further tailored by triggers due to ambiguities and uncertainties of the predictive value of an AI-model (like hallucination, opacity, or automation bias).

Therefore, the key question when using an AI system is in which mode the human uses it either task-oriented, supplementing their activities, or goal-oriented, replacing their activities to perform other tasks. This applies equally to autonomous vehicles and other autonomous systems, including AI systems.

A system failure occurs whenever the human uses the assistance system in goal-oriented mode and the system is operating in a design-exceeding mode, meaning it cannot perform its assistance function and will fail with P=1 (VDI-EE-AS 2021). Hence the decision of whether the personnel decision-maker acted in a task-oriented or goal-oriented manner, determines the risk assessment and would lead to completely different results regarding the system's safety and AI risk level.

For a comprehensive analysis of the overall AI system, all system functions can be represented as an event flow diagram (Figure 4). This allows for more nuanced risk assessments; for example, the importance of fact-checking

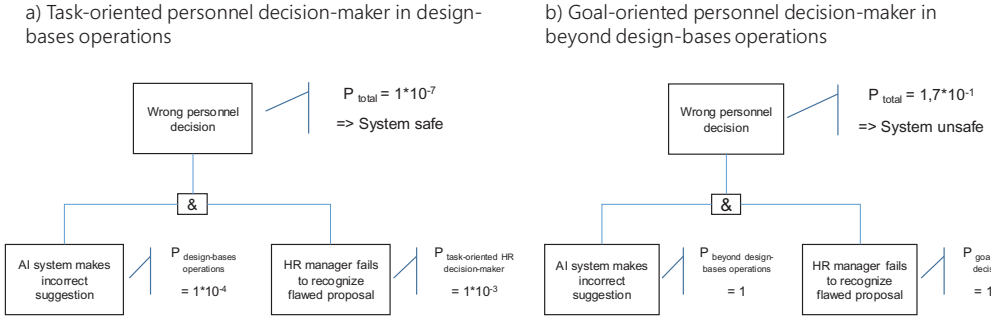

Figure 3: Fault tree of the system function "Data usage from the system structure of the AI system to support personnel selection".

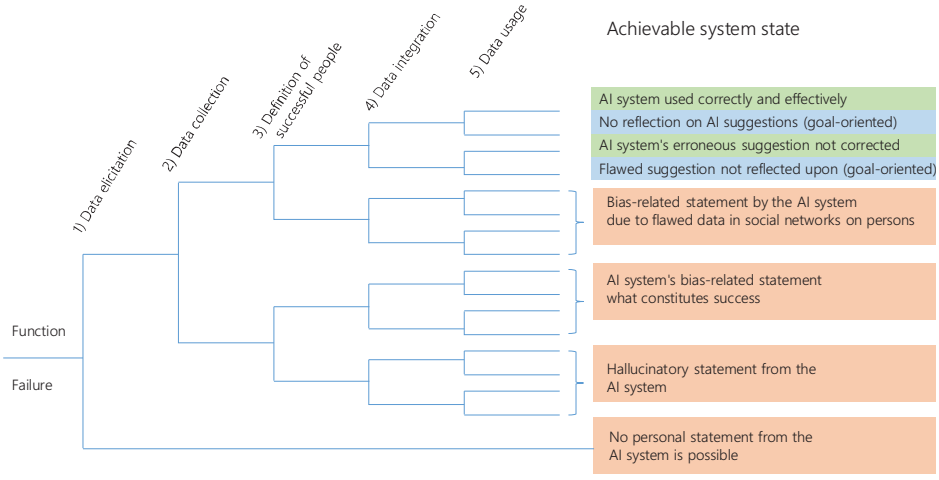

Figure 4: Flowchart of AI-based personnel decision-making.

during data collection (H2) or the valid definition of successful individuals (H3). The achievable system states can then be defined from an event flow diagram. Figure 4 shows a flow diagram of AI-based personnel decision-making. The detailed assessment of each safety-function again would need to determine task-oriented or goal-oriented risk contributions for these safety function in an analogue way to the above assessment.

Overall, there are only two success paths within the AI system (marked in green), which result from system function 5) and the fault tree according to Figure 3, provided all other system functions are successful. All other paths lead to inadequate system behavior; therefore, the AI system is unusable according to the AI act. While it could make statements about an applicant's suitability, these statements would not be reliable enough.

4. Outlook

VDI 4006 and the associated expert recommendations VDI EE and VDI EE VZ set important requirements for the reliability of an overall system. This paper demonstrates how modern systems, such as autonomous driving or artificial intelligence systems, can be assessed using VDI 4006.

It becomes clear that the scope of the overall system, as well as the correct assessment and differentiation of task-oriented and goal-oriented behavior, that these are critical for the accurate evaluation of the system's reliability. For the successful implementation of the EU Act on Artificial Intelligence Systems, a risk assessment of these systems according to VDI 4006 or similar is urgently required.

As Arenius & Straeter (2025) show, the effective use of operational data for system optimization according to VDI 4006 Part 3 is a good way to generate a sound data basis for safety management and to promote a safety culture (Straeter, 1997). Data analysis, operational management, and safety culture must be closely coordinated in this process (VDI-EE-VZ 2024).

This important distinction between task-oriented and goal-oriented behavior is also crucial for the increasing importance of the security sector in particular of AI IT systems in general. Methodologically speaking, security issues, as defined by VDI 4006, are always questions concerning goal-oriented human behavior; consequently, these issues can be analyzed and evaluated using the methodological framework of VDI 4006.

As the AI example presented demonstrates, system development is particularly critical for the

reliability of the overall system. For instance, the economically motivated decision to discontinue fact-checking within an AI system in the US states has a significant impact on the reliability of the achievable system states in its applicability.

The general relationship of Human Reliability and AI holds for both, "traditional" and generative AI-systems. Nevertheless, the goal oriented aspects will be even more crutial in generative AI-systems. Generative AI-systems change their functional spectrum to new, generative behavior and evolve post-deployment unspecified behaviour. They are hence comparable to Human-Human relationships. Accordingly they need similar concepts like Human-Human reliability design such as Team Ressource Mangement approaches (Helmreich & Sexton, 2002) and safety culture (Schein, 1992).

Economic and temporal considerations also play a major role in other current problem areas, such as supply and disposal. These areas are characterized primarily by their long lead times in time or space, and are therefore subject to ongoing consideration processes. Examples include the disposal of highly radioactive waste, with a project timeframe spanning several decades, or aspects of globalization (keyword: supply chain law). Here, cost-benefit analyses play a significant role. Corresponding assessment methods for human reliability are currently under development to evaluate such trade-offs regarding human reliability (see Fritsch 2025). These methods will be incorporated into future revisions of VDI 4006.

Ultimately, human reliability methods also contribute to the evaluation of AI systems themselves, because assessing the safety performance of AI systems using classical evaluation approaches and technical failure probabilities does not adequately reflect the adaptive nature of these systems. However, AI systems can be evaluated using human reliability methods – just like their biological counterparts (humans).

References

Apostolakis, G., Soares, C.G., Kondo, S. & Straeter, O. (2004, Ed.) Human Reliability Data Issues and Errors of Commission. Special Edition of the Reliability Engineering and System Safety. Elsevier.

Arenius, M. & Straeter, O. (2025). Organisatorisches Lernen durch die Berücksichtigung von Menschlicher Zuverlässigkeit im Eisenbahnsektor. In VDI-Wissensforum (Hrsg.), 32. VDI-Fachtagung Entwicklung und Betrieb zuverlässiger Produkte. Düsseldorf: VDI-Verlag. [Organizational learning through the consideration of human reliability in the railway sector. In VDI-Wissenforum (ed.), 32nd VDI Conference on the Development and Operation of Reliable Products. Düsseldorf: VDI-Verlag.]

Auto-Motor-Sport (2024) Tesla ist die tödlichste Automarke [Tesla is the deadliest car brand]. https://www.auto-motor-und-sport.de/verkehr/tesla-unfallstatistik-usa-toedlichste-automarke/

EU-Act (2025). https://www.europarl.europa.eu/news/de/headlines/society/ 20230601STO93804/ki-gesetz-erste-regulierung-der-kunstlichen-intelligenz

Fritsch, F. (2025). Selbstbewertung als Methode zur Steigerung der menschlichen Zuverlässigkeit. In VDI-Wissensforum (Hrsg.), 32. VDI-Fachtagung Entwicklung und Betrieb zuverlässiger Produkte. Düsseldorf: VDI-Verlag. [Self-assessment as a method for increasing human reliability. In VDI-Wissenforum (ed.), 32nd VDI Conference on the Development and Operation of Reliable Products. Düsseldorf: VDI-Verlag.]

Grabbe, N. (2024). The Contribution of Automated Driving to Road Traffic Safety – A Resilience Engineering Approach. Dissertation an der TU-München. https://mediatum.ub.tum.de/ doc/1720274/1720274.pdf

Helmreich, R.L. & Sexton, J.B. (2002). Managing threat and error to increase safety in medicine. Berliner Colloquium: Group Interaction in High Risk Environments. May 6th, 2002. Gottlieb Daimler and Karl Benz Foundation, 2002.

Hollnagel, E. (1999) Looking for Errors of Omission and Commission or the Hunting of the Snark revisited. Reliability Engineering and System Safety. Elsevier.

Hollnagel, E. (2012). FRAM, the Functional Resonance Analysis Method: Modelling Complex Socio-technical Systems. Ashgate Publishing Limited, England

Schein, E. H. (1992) Organizational Culture and Leadership, 2nd Edition. Jossey-Bass, San Francisco

Straeter, O. (2019) Human Reliability Assessment - State of the Art for Task- and Goal-related Behavior. In Varde, P.V., Prakash, R. V. & Joshi, N. S. (Eds.) Risk Based Technologies. Springer. Berlin, Heidelberg.

Straeter, O. (1997). Beurteilung der menschlichen Zuverlässigkeit auf der Basis von Betriebserfahrung. Köln: Gesellschaft für Anlagen- und Reaktorsicherheit. [translated: Sträter, O. (2000) Evaluation of Human Reliability on the Basis of Operational Experience. GRS-170. GRS.]Köln/Germany. (ISBN 3-931995-37-2)

Straeter, O. (Hrsg). (2019). Risikofaktor Mensch? – Zuverlässiges Handeln gestalten. Berlin: Beuth. [Human Factor as a Risk Factor? – Shaping Reliable Action. Berlin: Beuth.]

Sträter, O. (2000) Using the CAHR-method to derive cognitive error Mechanisms. In: Kondo, S. & Furuta, K. (eds) PSAM 5 – Probabilistic Safety and Management. Universal Academy Press. Tokyo, Japan.

Sträter, O. (2005). Cognition and safety - An Integrated Approach to Systems Design and Performance Assessment. Ashgate. Aldershot. (ISBN 0754643255).

Sträter, O. (2025). Erfordernisse an die menschliche Zuverlässigkeit und deren Abbildung in VDI Richtlinien und Expertenempfehlungen. In VDI-Wissensforum (Hrsg.), 32. VDI-Fachtagung Entwicklung und Betrieb zuverlässiger Produkte, VDI-Berichte 2448, S. 97-110. Düsseldorf: VDI-Verlag. [Requirements for human reliability and their representation in VDI guidelines and expert recommendations. In VDI Knowledge Forum (ed.), 32nd VDI Conference on the Development and Operation of Reliable Products, VDI Reports 2448, pp. 97-110. Düsseldorf: VDI Publishing.]

VDI 4006 (2026) - Menschliche Zuverlässigkeit – Teil 3: Menschliche Zuverlässigkeit - Methoden zur Ereignisanalyse. Beuth-Verlag. Berlin. [Human Reliability – Part 3: Human Reliability – Methods for Event Analysis. Beuth-Verlag. Berlin.]

VDI 4006 (2026) Menschliche Zuverlässigkeit – Teil 1: Ergonomische Forderungen und Methoden der Bewertung. Beuth-Verlag. Berlin. [Human reliability – Part 1: Ergonomic requirements and methods of assessment. Beuth-Verlag. Berlin.]

VDI 4006 (2026) Menschliche Zuverlässigkeit - Teil 2: Methoden zur quantitativen Bewertung menschlicher Zuverlässigkeit. Beuth-Verlag. Berlin. [Human Reliability - Part 2: Methods for the Quantitative Assessment of Human Reliability. Beuth-Verlag. Berlin.]

VDI-EE-AS (2021). Berücksichtigung menschlicher Zuverlässigkeit in der Gestaltung autonomer Systeme. VDI Expertenempfehlung. Berlin: Beuth. [Consideration of human reliability in the design of autonomous systems. VDI expert recommendation. Berlin: Beuth.]

VDI-EE-VZ (2024). Betriebliche Anforderungen zur Erreichung einer teamorientierten Arbeits- und Gesundheitsschutzkultur „Vision-Zero“. VDI Expertenempfehlung. Berlin: Beuth. [Operational requirements for achieving a team-oriented occupational health and safety culture “Vision-Zero”. VDI expert recommendation. Berlin: Beuth.]