Human Factors in Cybersecurity, Vol. 127, 2024, 63–73
https://doi.org/10.54941/ahfe1004766

Towards a Human-Centric AI Trustworthiness Risk Management Framework

Kitty Kioskli[1,2], Laura M. Bishop[3], Nineta Polemi[2,4], and Antonios Ramfos[5]

[1]University of Essex, School of Computer Science and Electronic Engineering, Institute of Analytics and Data Science (IADS), Colchester CO4 3SQ, United Kingdom
[2]Trustilio B.V., Vijzelstraat 68, Amsterdam 1017 HL, The Netherlands
[3]Airbus, The Quadrant, Celtic Springs Business Park, Newport NP10 8FZ, United Kingdom
[4]University of Piraeus, Department of Informatics, Piraeus 185 34, Greece
[5]Athens Technology Center, Athens, 152 33, Greece

ABSTRACT

The evolving landscape of Artificial Intelligence (AI) seeks to emulate human behaviour within socio-technical systems, emphasizing AI engineering to supplant human decision-making. However, an excessive focus on AI system autonomy raises concerns such as bias and ethical lapses, eroding trust and diminishing performance. Such a lack of human integration into the AI decision-making loop, may in turn leave organisations open to more cyber risk than its tools and techniques hope to mitigate. Efforts to address these challenges involve incorporating ethical considerations, leveraging tools like IBM's Fairness 360 and Google's What-If tool to enhance fairness. Trust in AI technology is complex, involving human acceptance, performance, and empowerment. Trustworthiness is scrutinized in relation to legal, moral, and ethical principles, aligning with human behavioural patterns and organizational responsibilities. The proposed framework integrates research from diverse disciplines to ensure the trustworthiness of AI-driven decision support systems, accommodating both the needs of human users and their own perceptions of trust. It extends the NIST AI Risk Management Framework by considering users' social attitudes and values as well as business objectives throughout the risk management cycle. The framework advocates co-creation and human experiment processes at all stages, fostering continuous trustworthiness improvement to establish 'trustworthy' AI systems that are ultimately and optimally by users.

Keywords: AI, Human factors, Trustworthiness, Risk management, Framework

INTRODUCTION

An Artificial Intelligence (AI) system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments (Russell & Norvig, 2020). Different AI systems vary in their levels of autonomy and adaptiveness after deployment (OECD).

© 2024. Published by AHFE Open Access. All rights reserved.

AI research is primarily focused on replicating and/or enhancing human behaviour within socio-technical systems. AI engineering endeavours to create systems that practically replace human decision-making, yet an overemphasis on machine autonomy can lead to biased and non-objective outcomes, making the systems susceptible to bias. Such distrust in AI systems not only results in potentially injurious decision-making but also diminishes user performance, autonomy, and job satisfaction. Should AI processes not fully represent an organisation's policies, people and culture, not only can it hamper the establishment of a safe environment, but challenges the ability to spot anomalies in output that may suggest security threat is present (Stevens, 2020). To address these issues, developers integrate ethical reflection processes, often engaging with ethicists, and employ technical tools like IBM's Fairness 360 and Google's What-If tool to enhance fairness in AI systems. However, achieving trustworthiness in AI technology goes beyond such technical measures, requiring a focus on empowering human users with design choices that grant them control and transparency over AI they need to remain productive, safe and secure (Floridi et al., 2018). Concerning the links between the properties of trustworthiness and the risks within the AI lifecycle, is published by UC Berkeley Centre for Long-Term Cybersecurity (Newman, 2023).

Trustworthiness in AI refers to the reliability, safety, transparency, fairness, and accountability of AI systems, ensuring they operate ethically, without bias, and in a manner that safeguards user privacy and rights (European Commission, 2019). The evaluation of trustworthiness should include human performance, satisfaction metrics, and an understanding of human decision support needs within the business context of the socio-technical system (Zhang et al., 2020). Thus, for an AI system to be trustworthy, it must align with the legal, moral, and ethical principles of its human users, considering organizational responsibility and liability in relation to business objectives (Jobin et al., 2019, AI Act).

In response to the societal imperative of democratizing AI amidst the digital transition and the impending wave of 'deep tech' innovation, the proposed framework introduces an innovative approach for assessing, and managing risks (technical and social) related to all dimensions of trustworthiness. The primary aims of this paper are: to adopt an AI risk management view engaging all users (e.g., developers, operators, integrators, etc.) of the AI system under assessment through interactive co-creation social experiments, interventions, and dialogues; to lead our adopted human-engaging approach to adequate, human-interpretable explanations of the AI system's decision-making process but also intelligently extract knowledge related to users' decision support needs, moral values, and the key business objectives of the AI system viewed as a socio-technical system. This dynamic interaction forms the basis for continuous trustworthiness improvement processes, where each cycle involves a human-centric assessment and the identification of corrective actions before the next improvement phase.

RISK-BASED METHODOLOGY FOR ENSURING TRUSTWORTHINESS

Before we propose the Trustworthiness Risk Management Framework (AI-TMF), our initial focus is on refining the definition of trustworthiness in AI hybrid decision support. This theoretical conceptual framework has been developed with the use of data found in the literature. It proposes suggestions on how data could potentially enhance and refine this framework. This includes the need for AI to be transparent in its reasoning, and adaptable and secure in its behaviour (Hou, 2021; Jacovi, 2021). Additionally, we explore the impact of the AI decision support system on the broader socio-technical system, considering it a significant factor in trustworthiness. A methodology for assessment, rooted in the Decision Intelligence discipline, will be developed. Concurrently, our multi-disciplinary research and co-creation practices aim to identify human characteristics influencing the perceived trustworthiness of an AI system.

AI systems are dynamic (main difference with ICT systems) with AI system lifecycle phases involving: i) 'design, data and models'; which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; ii) 'verification and validation'; iii) 'deployment'; and iv) 'operation and monitoring'. These phases often take place in an iterative manner and are not necessarily sequential. The decision to retire an AI system from operation may occur at any point during the operation and monitoring phase (OECD, ENISA 2022).

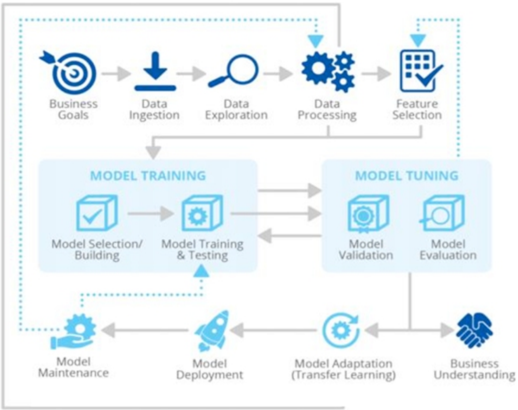

Figure 1: AI-lifecycle according to ENISA (ENISA, 2020).

The Human-Centric AI Trustworthiness Risk Management Framework is set to enhance the evaluation of trustworthiness in decision-supporting AI systems, introducing an innovative and human-centered methodological approach essential for assessing social and technical risks. This evaluation methodology takes a dynamic risk management perspective, aligning with established frameworks and standards (Fig. 2) in the entire lifecycle of the AI systems.

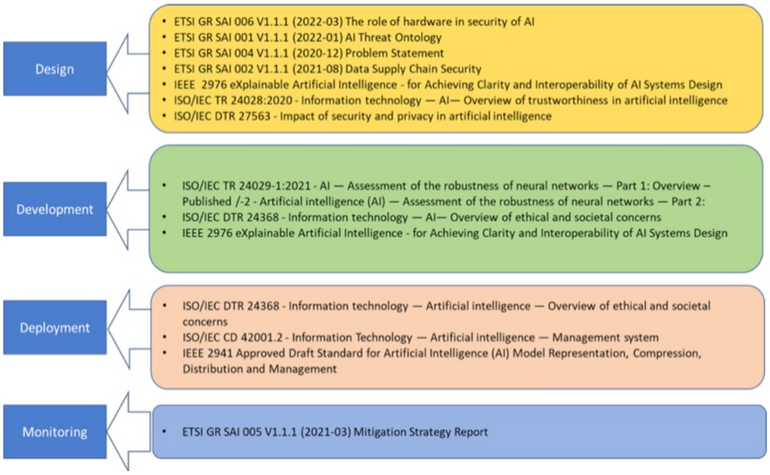
Design
• ETSI GR SAI 006 V1.1.1 (2022-03) The role of hardware in security of AI
• ETSI GR SAI 001 V1.1.1 (2022-01) AI Threat Ontology
• ETSI GR SAI 004 V1.1.1 (2020-12) Problem Statement
• ETSI GR SAI 002 V1.1.1 (2021-08) Data Supply Chain Security
• IEEE 2976 eXplainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design
• ISO/IEC TR 24028:2020 - Information technology — AI— Overview of trustworthiness in artificial intelligence
• ISO/IEC DTR 27563 - Impact of security and privacy in artificial intelligence
Development
• ISO/IEC TR 24029-1:2021 - AI — Assessment of the robustness of neural networks — Part 1: Overview – Published /-2 - Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2:
• ISO/IEC DTR 24368 - Information technology — AI— Overview of ethical and societal concerns
• IEEE 2976 eXplainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design
Deployment
• ISO/IEC DTR 24368 - Information technology — Artificial intelligence — Overview of ethical and societal concerns
• ISO/IEC CD 42001.2 - Information Technology — Artificial intelligence — Management system
• IEEE 2941 Approved Draft Standard for Artificial Intelligence (AI) Model Representation, Compression, Distribution and Management
Monitoring
• ETSI GR SAI 005 V1.1.1 (2021-03) Mitigation Strategy Report

Figure 2: Guidelines associated with the lifecycle implementation of AI systems (ENISA, 2023).

The envisaged trustworthiness evaluation methodology shall go beyond the conventional approach by incorporating considerations for vulnerabilities related to fairness, technical accuracy, robustness, and adherence to the EU legal framework for trusted AI. What sets the trustworthiness evaluation methodology apart is its novel integration of the human perspective and a broader socio-technical systems outlook in the risk management-based assessment of trustworthiness. Helping predict not just system behaviour, but human behaviour and how they co-exist.

THE AI TRUSTWORTHINESS MANAGEMENT FRAMEWORK (AI-TMF)

Our proposed AI-TMF is dynamic theoretical framework that empowers human recipients of AI-based decision support by enabling:

- A dynamic risk management perspective, aligning with established frameworks such as the AI Risk Management Framework proposed by NIST (NIST_AI_RMF), NIST Special Publication 1270, risk management standards (e.g., ISO31000, ISO27001, ISO27005), and specific standards within the AI systems lifecycle according to ENISA 2022 (refer to Figure 2) in order to estimate the technical driven risks.
- Effective elicitation of users' decision support needs, moral values, and key success factors based on human psychological, ethical, and behavioural analysis, along with advanced organizational and decision theory approaches. Social co-design experiments, behavioural interventions, and dialogs with the users (administrators, developers, operators etc) of the AI system under assessment are the instruments used to measure the social driven risks.
- Human-centric evaluation (i.e., usability testing) and optimization of trustworthiness in terms of fairness, technical accuracy, and robustness, implementing dynamic, inherently transparent risk-assessment approaches.

- Enhanced explainability regarding the trustworthiness of AI-based decision support through anomaly detection indicators related to fairness, technical accuracy, robustness, and the socio-technical environment.
- Human-centric actions for optimizing trustworthiness based on risk mitigation approaches concerning fairness, technical accuracy, robustness, and the socio-technical environment. Mitigation actions include not only technical controls but also non-technical mitigation actions including awareness, behaviour change interventions, and trainings.

The AI-TMF framework diligently executes a series of steps (risk management based) outlined in Figure 3 for every AI system implementation cycle. This iterative process ensures ongoing improvements, adaptation, and alignment with evolving socio-technical dynamics, reinforcing the commitment to establishing and maintaining appropriate trustworthiness in AI systems to maintain use and avoid abuse or misuse. Each step adopts technical driven Risk Assessment methodologies (e.g., NIST, ENISA to estimate the technical risks and social experiments, behavioural interventions and dialogs are used to estimate the social risks in the social environment that the AI systems under assessment operate (Figure 3).

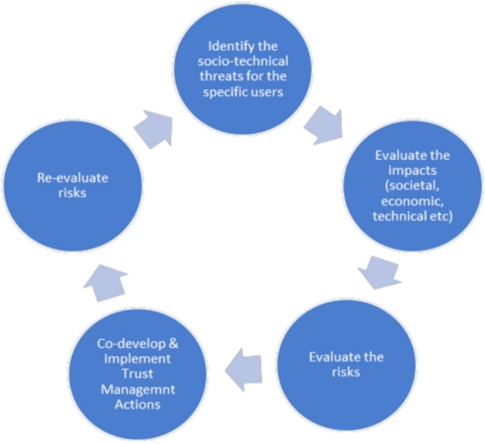
Identify the socio-technical threats for the specific users
Evaluate the impacts (societal, economic, technical etc)
Evaluate the risks
Co-develop & Implement Trust Managemnt Actions
Re-evaluate risks

Figure 3: The AI trustworthiness management framework (AI-TMF).

Recognizing the pivotal role of the human perspective, the framework engages in research to implement intelligent human-AI dialogues to estimate the social-driven risks. This involves providing comprehensive explanations for AI system decisions, including details on the employed AI approaches, models, algorithms, and training data sets in a language appropriate to the people, culture and objectives of the organisation. Simultaneously, the framework aims to optimize human trust in the AI system, aligning with user expectations and values. This approach is pivotal for establishing transparency and fostering confidence in AI decision-making processes as well as aid detection of anomalies and potential threat.

THE AI-TFM CONCEPTUAL ARCHITECTURE

The proposed AI-TFM conceptual architecture, is illustrated in Figure 4 where we provide detailed design considerations for each module within the conceptual architecture. For example, the notion of fairness in AI encompasses concerns for equality and equity, addressing issues such as bias and discrimination. The framework recognizes that perceptions such as fairness vary among AI users and may shift depending on the system's usage. For instance, even AI systems deemed non-biased may lack fairness, as demonstrated when training data sets eliminate demographic bias but remain inaccessible to disabled individuals. Bias can be introduced at any phase of the AI system implementation lifecycle, from design to validation. The framework assesses fairness using indicators defined by NIST, covering systemic bias, computational bias, and human bias. The 'AI-based Assessment of Fairness' module computes metrics for these three major categories of AI bias, employing the principles of an AI-driven anomaly detection system.

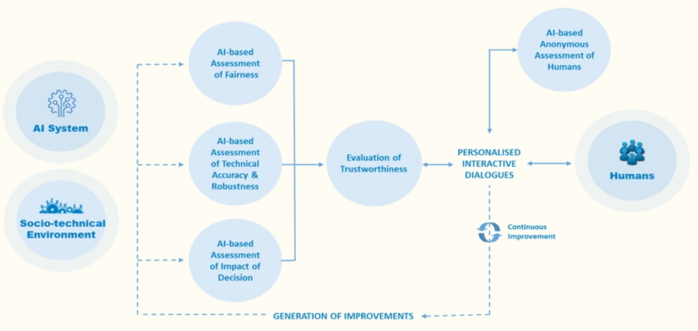
AI System
Socio-technical Environment
AI-based Assessment of Fairness
AI-based Assessment of Technical Accuracy & Robustness
AI-based Assessment of Impact of Decision
Evaluation of Trustworthiness
AI-based Anonymous Assessment of Humans
PERSONALISED INTERACTIVE DIALOGUES
Humans
Continuous Improvement
GENERATION OF IMPROVEMENTS

Figure 4: Socio-technical approach in the AI-TMF.

Co-creation sessions with citizens and stakeholders would ensure that the 'AI-based Assessment of Fairness' module meets their functional expectations and aids in creating initial training data sets for the developed AI models. Co-creation practices involve fostering collaboration among diverse stakeholders to collectively identify, assess, and mitigate risks associated with a system or process. This includes conducting cross-functional workshops, facilitating threat modelling sessions, and engaging in iterative risk assessments with continuous feedback loops. Joint scenario planning allows teams to simulate and evaluate responses to potential risks, refining strategies based on collective insights. Establishing a security champions program and maintaining collaborative documentation further ensures a culture of shared responsibility and transparency. Through these co-creation practices, organizations can tap into the collective expertise of their teams, proactively addressing potential vulnerabilities and embedding trustworthiness into the core of their systems.

The output of this module contributes to the 'Evaluation of Trustworthiness' module, as depicted in Figure 4. The assessment of technical accuracy and robustness in the 'AI-based Assessment of Technical Accuracy & Robustness' module is implemented based on the principles of an AI-driven anomaly detection system. Reliability and validity, understood in terms of accuracy and robustness, may be assessed through audits or monitoring. The framework leverages AI for assessing accuracy and robustness through activation analysis, examining patterns in deep neural network classifiers. Activation analysis serves as an evaluation of both robustness and accuracy, complementing other approaches such as Bayesian neural networks. Additionally, it facilitates enriched explanations beneficial to domain experts involved in the development and use of hybrid decision systems. Co-creation sessions model relevant user, task, and contextual characteristics, aligning them with the required levels of robustness and accuracy detailed through activation pattern analysis in deep neural networks. This alignment enables sliding decision-making, and assessments for accuracy and robustness in human-AI decision-making feeding into the 'Evaluation of Trustworthiness' module, allowing for a dynamic assignment of decision-making agency between human and machine actors in the AI system.

To include the impact of decisions on the relevant socio-technical system in the trustworthiness optimization process, an "AI-based Assessment of Impact of Decision" module will be constructed. First, the human user will use the developed Decision Intelligence (DI) methodology (Pratt et al., 2023) and supporting graphics to create a qualitative model of the socio-technical environment, considering possible actions, key performance indicators (KPIs), external factors, and their pairwise relations. Second, using co-created data-driven approaches, a DI AI simulation will be trained to generate a quantitative simulated model of the socio-technical system, potentially learning pairwise interactions quantitatively from co-created data sets. The simulated model provides scenarios related to the AI system's decision support and the projected effects on the KPIs of the overall socio-technical system, contributing to the 'Evaluation of Trustworthiness' module. The socio-technical environment model and AI-based simulation inherently bring risks, necessitating iterative trustworthiness assessment by the user. Evaluation of trust in the DI simulation results is crucial, and user feedback highlighting issues in the DI model or simulation is addressed by correcting the socio-technical environment model or the generated AI-based DI simulation. This user feedback is gathered by "Personalized Interactive Dialogues" and acted upon in the "Generation of Improvements," as shown in Figure 4.

The 'AI-based Anonymous Assessment of Human' module aims to enhance the evaluation of trustworthiness by incorporating the human perspective and assessing both behavioural and moral value aspects. Whilst a succinct definition of trust is still to be determined, it is noted that an artefact is considered trustworthy should it be perceived as (a) capable, (b) honest, and (c) kind (Gefen et al., 2008; Viklund, 2002). With such views believed to be influenced by not just the output of the artefact, but end-user individual differences such as personality, perceptions and beliefs (e.g., Riedl, 2022; Sharan et Romano, 2020). An analysis of these aspects at both the individual and cultural levels in relation to AI are therefore imperative, should a universal picture be painted in relation to trustworthiness across both technical and human layers. For this particular aspect of the framework, we draw upon investigative psychology research and behavioral science utilising frameworks such as Fogg's behavioral model (Fogg, 2009), that posits that the likelihood of a behavior occurring is influenced by Motivation (M), Ability (A), and an appropriate Trigger (T), as depicted in Figure 5. In addition, consideration will be given to the Unified Theory of Acceptance and Use of Technology Model (UTAUT) that is now shifting focus towards the adoption and use of AI tools and techniques with trust an important factor (Venkatest, 2022).

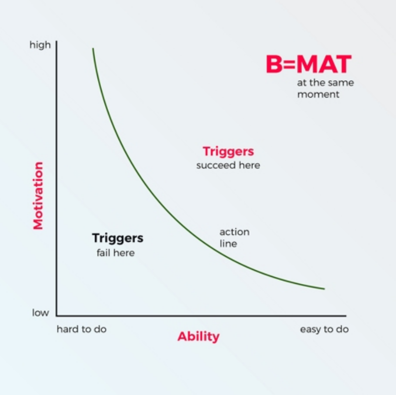

Figure 5: Fogg's model, 2009 (Ray, 2021).

Human profiles can be proposed based on five distinct categories of traits with specific attributes and measurement scales: personality, social-behavioural factors, technical awareness and efficacy, motivation, and an appropriate trigger. Notably, studies have shown that people tend to trust inhuman-generated profiles more than AI-generated ones (Kioskli & Polemi, 2022). In our framework, we share this perspective and plan co-creation sessions, including workshops and living labs with stakeholders and pilot users, to build user categories and capture their profiles. Anonymous protocols and mechanisms will be employed during development, storage, and sharing phases to ensure the anonymity of the profiling process. The moral value evaluation of the human factor within the framework will involve a critical exploration of moral judgment and a self-reflective dialogical awareness. Specifically, the development of the moral values evaluation will consider applied ethical theories (Kim et al., 2019), essential for motivating agents to behave morally. Not least, investigations will consider the impacts such trust profiles may have on the abuse, misuse and disuse of AI tools and techniques to aid vulnerability identification.

TRUSTWORTHINESS ASSESSMENT USING AI

Our approach to assessing trustworthiness revolves around the fundamental concept of risk assessment accompanied with social co-creation experiments, interventions, living labs and dialogs. This choice is rooted in the acknowledgment that the majority of human decisions inherently involve varying degrees of risk evaluation. Whether done consciously and explicitly or unconsciously, decision-makers routinely engage in weighing the pros and cons associated with their choices. The implementation of the 'AI-based Evaluation of Trustworthiness' module is anchored in well-established Risk Assessment (RA) frameworks. The module generates outputs that feed into the system-under-test model, manifesting as distinct vulnerabilities that trigger an automatic recalculation of risks. The recalculated risks yield altered risk levels and propose controls to mitigate any elevated risk levels identified. Human operators can select based on the business objectives, use of the systems, technical and social attributes and apply these controls to the actual system (technical controls) and its users, (social controls will also be included e.g., awareness, behaviour change interventions) thereby reducing the likelihood of increased unfairness or diminished accuracy.

Our deliberate choice of a knowledge-based approach enhances transparency, a critical factor in achieving advanced explainability levels for any AI decision support provided to humans. Previous work in the Systems Trust Model (STM) has led to the development of a "threat path explorer," allowing users to navigate from identified risks to their root causes, exploring the impact of controls on systemic risk (Phillips et al., 2022). While AI and machine learning tools may drive the STM, the path from events like heightened unfairness to risks affecting trustworthiness remains clear and evident to the user.

In particular, our methodology involves a step-by-step co-design process that offers comprehensive guidance for developing user-centric interventions. We provide a detailed outline for modifying these interventions into practical and acceptable prototype behavioral control mechanisms. Engaging users in this process allows them to contribute suggestions and propose solutions, ensuring greater acceptance within the broader target audience. This user-centric approach aims to foster a collaborative environment for refining and optimizing trustworthiness interventions in AI decision support systems, in turn improving not only system productivity, but its safety and security.

CONCLUSION

In conclusion, our research underscores the significance of incorporating an AI risk assessment-based approach enriched with social experiments, behavioural interventions, and dialogs in evaluating the trustworthiness of AI systems. By aligning our proposed AI-TMF with existing Risk Assessment (RA) frameworks, we leverage the well-established principles of risk analysis to identify vulnerabilities and recalibrate technical and social risks within the system-under-test. The automatic recalculation process, coupled with suggested controls, empowers human operators to intervene effectively, mitigating potential risks such as unfairness or accuracy issues. This knowledge-based approach not only prioritizes transparency but also ensures an advanced level of explainability, crucial for building user trust in AI decision support. Our work within the Systems Trust Model (STM) framework has resulted in practical tools like the "threat path explorer," providing users with a clear understanding of the causal relationships between identified risks and potential threats, further enhancing the user's ability to manage and improve system trustworthiness.

Furthermore, our proposed AI-TMF enriches the existing RA methodologies with social experiments. AI-TFM introduces a step-by-step co-design process that promotes user-centric interventions for enhancing trustworthiness assessments estimating not only technical but also social risks. By involving users in the development and modification of behavioral interventions, we tap into their insights and preferences, ensuring a more tailored and widely accepted approach. This collaborative model fosters a dynamic feedback loop, empowering users to make suggestions and propose solutions that resonate with the broader target audience. Our continuous research aims to advance the understanding in assessing the social and technical dimensions of trustworthiness in AI systems, paving the way for more responsible and user-friendly decision support solutions serving the people and their democratic, ethical values and morals.

ACKNOWLEDGMENT

The authors would like to acknowledge the 'Human-centered Trustworthiness Optimisation in Hybrid Decision Support' (THEMIS 5.0) project under grant agreement No. 101121042 and the 'Fostering Artificial Intelligence Trust for Humans towards the optimization of trustworthiness through large-scale pilots in critical domains' (FAITH) project under grant agreement No. 101135932. The views expressed in this paper represent only the views of the authors.

REFERENCES

AI Act: https://eur-lex.europa.eu/legalcontent/EN/TXT/DOC/?uri=CELEX:52021PC0206.

Benson, H. (2000). Socratic Wisdom. Oxford: Oxford University Press.

ENISA (2020) https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges.

ENISA (2023) https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V. & Valeriani, A. (2018). AI4 People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707.

High-Level Expert Group on Artificial Intelligence - European Commission. (2019). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Hou, Y. et al. (2021). Artificial Intelligence is a promising prospect for the detection of prostate cancer extracapsular extension with MP-MRI: A Two-center comparative study [Preprint]. doi: 10.21203/rs.3.rs-298296/v1.

Jacovi, A. et al. (2021). 'Formalizing trust in artificial intelligence', Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency [Preprint]. doi: 10.1145/3442188.3445923.

Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1(9), 389–399.

Kim, T. W. and Mejia, S. (2019). "From Artificial Intelligence to Artificial Wisdom: What Socrates Teaches Us," in Computer, vol. 52, no. 10, pp. 70–74.

Kioskli K., & Polemi N. Estimating attackers' profiles results in more realistic vulnerability severity scores. Proceedings of the AHFE2022, July 24–July 28, 2022, New York, New York, USA, 53 (1), 138–150. Springer, Elsevier, CRC.

Newman, J. (2023). A taxonomy of trustworthiness for Artificial Intelligence. https://cltc.berkeley.edu/wpcontent/uploads/2023/12/Taxonomy_of_AI_Trustworthiness_tables.pdf

OECD: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

Phillips, Taylor, et al. (2022). System Security Modeller. Zenodo.

Pratt, L., Bisson, C. and Warin, T. (2023). 'Bringing advanced technology to strategic decision-making: The Decision Intelligence/Data Science (DI/DS) integration framework', Futures, 152, p. 103217. doi: 10.1016/j.futures.2023.103217.

Ray, D. (2021). https://bootcamp.uxdesign.cc/foggs-behaviour-model-a-framework-for behaviour-change-fd6ce4b0a1f2

Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Stevens, R. et al. (2020). AI for science: Report on the Department of Energy (DOE) town halls on artificial intelligence (AI) for Science [Preprint].

Zhang, B., Liao, Q., & Zhu, T. (2020). Explainable AI: A Survey of Black Box Deep Learning Models Interpretability. IEEE Access, 8, 17958–17975.