A Multidisciplinary Approach to Trustworthy AI in the Digital Society

Electra MITAN, Delia NEACȘU, Silvia OVREIU, Daniel SAVU
National Institute for Research and Development in Informatics – ICI Bucharest
electra.mitan@ici.ro, delia.radulescu@ici.ro, silvia.ovreiu@ici.ro, daniel.savu@ici.ro

Abstract: This review paper presents a multidisciplinary examination of trustworthy Artificial Intelligence (AI) in the digital society. It integrates technical, ethical, legal, and social perspectives to analyze key principles such as safety, transparency, fairness, and accountability. The paper discusses regulatory frameworks including the European regulation and international standards, alongside practical applications in healthcare, finance, education, public administration and justice, industry, and cybersecurity with comparative analysis. Challenges in AI governance, bias mitigation, and human oversight are addressed. The paper proposes an integrated framework for developing and evaluating trustworthy AI systems, emphasizing the multidisciplinary collaboration and audit for AI adoption.
Keywords: AI Systems, Trustworthy AI, Regulation, Governance, Responsible AI Systems.

INTRODUCTION

AI has been gradually integrated into almost all areas of the society as an essential technology for digital transformation. It optimizes workflows, reduces costs, and drives innovation, as can be seen in disease diagnosis, industrial process automation, personalized online recommendations, and financial decision-making. In Romania, large companies such as BCR, ING or Raiffeisen use AI solutions for credit scoring or virtual assistance, while companies like UiPath or start-ups like Druid AI and TypingDNA developed advanced technologies for automation, biometric recognition, and digital security. Large public administrations have begun to use AI in traffic management or urban data analysis, and in the field of robotics, centers such as RoboHub are exploring AI applications in human-machine interaction and in the creation of collaborative robots. However, as AI spreads, major risks also emerge, often insufficiently understood or regulated. Among the most worrying are the violation of privacy, through the massive and non-transparent collection of personal data, algorithmic discrimination, when models reproduce the bias in the training data, and the lack of responsibility in the automated decisions with a direct impact on people's lives.

For example, AI systems can automatically reject credit applications based on social factors, without any clear justification. Or they can make decisions in recruitment or justice, in a non-transparent way, with potentially unfair effects. The generative AI models can create

https://doi.org/10.54851/v7i2y202504

false content - deepfakes, fake news - that can be used for manipulation or blackmail, seriously affecting the public trust and democratic processes. These risks are not theoretical.

For example, in Romania, the algorithms used in the fund allocation processes, social assistance or school admission processes have been proved to be non-transparent and difficult to control. Without an effective human intervention, such systems can perpetuate inequalities and affect fundamental rights, especially for vulnerable groups. The lack of legal regulations, audit standards, and mechanisms for tracing the decision-making process further complicates the situation. In this complex technological landscape, human oversight is poorly defined and the responsibility for automated decisions is not assumed. Since 2024, the AI Act (AI Act, 2024) has defined the concept of *„trustworthy AI"* and required compliance with fundamental values: transparency, robustness, fairness, accountability, and human-centeredness.

Under this law, AI systems are classified according to the risk they pose, from low risk to unacceptable risk. The law imposes strict requirements on systems with a high impact. Thus, audit, continuous human oversight, data protection, and decision traceability have become mandatory. In this framework, with a multidisciplinary approach, technological, legal, social, and ethical expertise may be combined. The systems may accomplish several characteristics: they must be technically efficient, explainable, fair, and secure (Amann et al., 2020). There is a need for auditable algorithms, clear legal frameworks, inclusive public policies, and a strong ethical culture around the development and use of AI. Through standardization and regulation, and through transdisciplinary education, the specialists can be trained to design, implement, and govern AI systems that respect the fundamental rights and freedoms (Herzog, 2025).

This review paper aims to explore the concept of trustworthy AI from an integrated perspective, presenting a conceptual framework for developing trustworthy AI, the technical, ethical, legal, and social dimensions of trust, concrete examples of responsible application in various fields with comparisons, as well as future directions and major challenges of effective governance in the digital society. The emphasis is on the need for a balance between technological innovation and the protection of fundamental human values.

This paper differs from the existing Organisation for Economic Co-operation and Development (OECD) and European Union (EU) reports on trustworthy AI primarily through its integrated multidisciplinary approach. While OECD and EU documents, such as the OECD AI Principles and AI Act, provide foundational ethical guidelines, regulatory frameworks, and risk classifications for AI, this paper goes further by combining technical, ethical, legal, social, and psychological perspectives to analyze the trustworthy AI holistically. The paper analyzes concrete examples of AI implementations across different sectors - healthcare (e.g., Kepler Night Nurse AI), finance (Qualco), education (HireVue), justice, public administration, industry, and cybersecurity - to illustrate practical applications of the trust principles. The importance of a transdisciplinary collaboration among developers, ethicists, legal experts, policymakers, and civil society for effective AI governance is discussed.

The paper proposes an integrated framework for developing and evaluating trustworthy AI systems, emphasizing the multidisciplinary collaboration and audit for AI adoption.

The remainder of the paper is structured as follows: Section 2 presents the *Recent research on multidisciplinary approach of trustworthy AI*; Section 3 analyses the *Foundations of trustworthy AI* from technical, ethical, legal, and social dimensions standpoint: important characteristics of trustworthy AI; associated risks and the need for regulation of AI; fundamental dimensions of trustworthy AI; technical issues and ethical basis of trust in AI; a framework for developing a trustworthy AI system with a multidisciplinary approach. *Dimensions of trust from a multidisciplinary approach* are presented in section 4. In section 5, *Standardization initiatives and AI*

governance are described. The next section presents *Examples of trusted AI applications* that integrate applicable standards and trust dimensions, with comparisons made across various sectors: health, finance, education and human resources, public administration and justice, industry, and cybersecurity. *Challenges and future directions in developing trustworthy AI* are described in section 7. Conclusions are given in the last section.

RECENT RESEARCH ON MULTIDISCIPLINARY APPROACH OF TRUSTWORTHY AI

The rapid integration of AI into the digital society has amplified the need for trustworthy AI systems. This trustworthiness requires a multidisciplinary approach, drawing on technical, ethical, legal, and social sciences (Kaur et al., 2022; Al-Kfairy et al., 2024; Sharma, 2024; Polemi et al., 2024; Bareis, 2024, Rădulescu et al., 2025). Studies emphasize that the trustworthy AI should be legal, ethical, and robust (Rodríguez et al., 2023; Kaur et al., 2022). The collaboration of technologists, policymakers, and end-users is necessary for developing governance models, regulatory frameworks, and practical AI tools (Sharma, 2024; Polemi et al., 2024; Bareis, 2024). Some of the practical strategies for trustworthy AI are:

- auditing and certification processes to ensure compliance and accountability (Rodríguez et al., 2023; Liu et al., 2021; Sharma, 2024);
- explainable AI (XAI) frameworks to enhance transparency and user trust (Finzel, 2025; Chamola et al., 2023; Malamuthu et al., 2025);
- education, communication, and training for users and developers (Polemi et al., 2024; Kusters et al., 2020; Li et al., 2024);
- adaptive governance models that evolve with technological and societal changes (Sharma, 2024; Bareis, 2024; Ahmed et al., 2025).

A comparison of the studies on multidisciplinary approaches to the trustworthy AI is done in Table 1.

Table 1. *A comparison of studies in multidisciplinary approaches to trustworthy AI*

Paper	Methodology	Focus Area	Results
(Rodríguez et al., 2023)	Theoretical and regulatory analysis	Multidisciplinary frameworks, regulation	Identifies seven requirements and emphasizes auditing and regulation for responsible AI
(Li et al., 2021)	Review	Lifecycle approach, technical and societal trust	Proposes a framework for trustworthiness across AI lifecycle
(Kaur et al., 2022)	Survey	Requirements, validation, and standardization	Analyzes fairness, explainability, accountability, and reliability in trustworthy AI
(Thiebes et al., 2020)	Conceptual framework	Foundational principles, data-driven research	Proposes five foundational principles and a research framework for AI
(Sharma, 2024)	Literature synthesis	Multistakeholder governance	Advocates for multi-stakeholder models for AI adoption

Trustworthy AI involves balancing multiple, often conflicting, criteria such as transparency, fairness, robustness, privacy, accountability, and human oversight. This complexity aligns well with the structure of multi-criteria decision problems, where various criteria must be evaluated and weighted to reach an optimal or compromise solution (Alsalem et al., 2024; Mitan, 2022; Ali et al., 2023; Rădulescu and Rădulescu, 2024; Voronin and Savchenko, 2024).

While earlier papers such as Rodríguez et al. (2023) and Li et al. (2021) provide theoretical

and regulatory analyses or propose lifecycle-spanning frameworks for trustworthy AI, the present paper integrates these perspectives with a multidisciplinary approach. It combines technical, ethical, legal, social, and psychological dimensions into a single framework that is grounded in practical governance.

Unlike Kaur et al. (2022) and Thiebes et al. (2020), which focus on requirements, validation, standardization or conceptual foundational principles, this paper goes further offering a framework for developing trustworthy AI, including the impact assessment, design, audit, and monitoring steps that are aligned with global standards (ISO, IEEE) and legal frameworks (the EU AI Act). The paper presents sector-specific examples from healthcare, finance, education, justice, industry, and cybersecurity with comparisons of trustworthy AI strengths, weaknesses, and outcomes.

Compared to Sharma (2024), who advocates a multi-stakeholder governance, this paper situates such governance within a framework and aligns it directly with technical and regulatory tools, moving from advocacy to implementation.

The paper advances the state of trustworthy AI research by bridging theoretical frameworks with organizational, technical, ethical, and social strategies, providing a practically deployable roadmap rather than abstract principles.

FOUNDATIONS OF TRUSTWORTHY AI: TECHNICAL, ETHICAL, LEGAL, AND SOCIAL DIMENSIONS

AI has become a central technology that is profoundly influencing the way people live, work, and interact in the digital society. The use of AI systems brings significant benefits, such as increased efficiency, easy access to information, and personalized services, but also raises numerous challenges related to security, privacy, fairness, and responsibility. In this context, the concept of trustworthy AI becomes an imperative necessity, defining those AI systems that are safe, transparent, fair, and developed responsibly, respecting fundamental rights and human values.

Important characteristics of trustworthy AI

A trustworthy AI system must meet the following fundamental criteria: safety and robustness, transparency and explainability, fairness and non-discrimination, responsibility and protection of personal data.

- Safety and robustness refer to the system's ability to function consistently and reliably in the face of variations, disruptions or even deliberate attacks (adversarial attacks), thus avoiding errors or unwanted behaviors with a negative impact on human and social values.
- Transparency implies the availability of relevant information on the functioning of the algorithm, accessible in a way that allows the understanding, evaluation, and auditing of the automated decisions to confirm their compliance with the social and ethical norms. Related to transparency is the concept of explainability, which consists of the system's ability to provide intelligible and motivated decisions to users and authorities.
- Fairness is essential to prevent algorithm-induced prejudice and discrimination. This criterion involves eliminating any algorithmic bias and ensuring an equal and non-discriminatory treatment for all users, regardless of their demographic or socio-economic characteristics.
- Responsibility requires the existence of mechanisms through which the responsibility for AI decisions and actions can be assumed, so that any errors or abuses can be effectively remedied.
- Last but not least, respecting confidentiality and protecting personal data are indispensable for protecting fundamental rights such as dignity, privacy, freedom, and equality.

Associated risks and the need for regulation

In addition to the benefits offered, the use of AI brings into question a series of major risks, including: the risk of disinformation and manipulation of users by generating

false content (fake news, deep fakes) that is difficult to detect, seriously affecting the public image and trust in the democratic processes; the loss of human control over automated systems, which can lead to abuses or errors with significant consequences; the lack of traceability and responsibility, which makes it difficult to understand how a decision was made and, in the absence of clear standards, impossible to apply an effective control and assign responsibility. To control these risks and ensure the ethical and safe operation of AI, the EU has proposed a classification of AI systems into four risk levels, each imposing different requirements and restrictions, as detailed in Table 2:

Table 2. *Classification of the risk level of AI systems*

Risk level	Exemples	Applicable regime
Unacceptable	Real-time facial recognition systems in public spaces without authorization, manipulative AI	Completely prohibited Systems that pose unacceptable risks to fundamental rights and the safety of individuals are completely prohibited
High	AI systems used in: health, transportation, education, justice, security, which require rigorous evaluation	Strict obligations: testing, documentation, human supervision; rigorous auditing. Requires CE marking before placing on the market
Limited	Chatbots, recommendation systems	Transparency requirements; providing clear information to the end user about functionalities and limitations
Low	Games, spam filters	No specific regulations; facilitates innovation and widespread use of low-impact AI technologies

Fundamental dimensions of trustworthy AI

The development and implementation of trustworthy AI requires a multidisciplinary approach that integrates the following eight dimensions (Vincent-Lancrin & van der Vlies, 2020) (see Table 3):

Table 3. *The main dimensions of trust in AI*

Dimension	Description	Key Aspects	Measurement method	Implementation example
Human supervision	AI systems must allow human intervention (to control or stop the system in critical situations), ensuring the decision-making autonomy of the people involved	Respect for human autonomy, continuous involvement in all phases of the AI cycle	Intervention log, audit	Human-in-the-loop models, real-time monitoring of automated decisions

Technical robustness and safety	AI systems must be resilient to unforeseen conditions, avoid dangerous errors, and maintain their performance even in the face of cyberattacks or disrupted data	Reliable operation, resistance to errors, attacks, disruptions, and predictable results	Resilience tests, stability indicators	Stress testing, anomaly detection, redundancy in critical processes
Privacy and data governance	AI systems ensure personal data management according to legal regulations, and protect user privacy, being transparent in data collecting, storing, and processing	Personal data protection, responsible management, and transparency regarding data use	Data access audit, data protection impact assessments	Secure storage, federated learning (without centralizing sensitive data)
Equity and non-discrimination	AI systems identify prejudices in data and algorithms to prevent discrimination based on gender, ethnicity, age or other characteristics	Reducing and auditing bias, fair and inclusive treatment for all social groups	Equity indicators	Ethical auditing, testing with balanced datasets, analysis of results
Transparency and explainability	AI systems must provide understandable explanations of automated decisions for users and for control bodies, supporting audits and challenging decisions	Clarity of operation and decisions, access to algorithm reasoning for users and auditors	Transparency audit, user comprehension level	Interfaces with visual explanations, public reports on model performance
Responsibility and auditability	AI developers and users must take accountability for automated effects, including the possibility of external audit and review of wrong automated decisions	Traceability, assuming the consequences of decisions, and mechanisms for redressing damages	Number of findings remedied, degree of ethical compliance	Ethical governance, automated audit platforms

Social impact and well-being	The social impact of AI is continuously assessed to adapt technologies to support sustainable development and avoid unexpected negative consequences	Monitoring the socio-economic, cultural, and environmental impact, and compliance with local values	Public perception surveys, social impact indicators	Periodic ethical assessments, responsible public-private partnerships
Inclusion and diversity	Ensuring diverse social representation and involvement in AI decision-making process responds to communities' needs & reduces the risks of marginalization / exclusion	Active involvement of vulnerable groups and diverse stakeholders in the development and monitoring	Diversity indicators, accessibility assessments	Inclusive design, public consultations

The technical and ethical basis of trust in AI

At a technical level, a basis for trust in AI is explainable and robust algorithms that can perform correctly under diverse conditions and provide decisions that are accessible to users and auditors. XAI reduces the perception of a "black box" and allows the verification of the system behavior, thereby contributing to the transparency and decision-making responsibility, strengthening the cognitive trust, based on evidence and performance (Kyriakou and Otterbacher, 2023). From an ethical point of view, AI must respect fundamental human rights and values (Díaz-Rodríguez, 2023). Fairness stems from the Kantian philosophy of respect for the person as an end in itself, the implication of non-discrimination, and moral responsibility for the consequences of action. Other relevant philosophical perspectives include the utilitarianism (maximizing the common good and reducing injustices), the virtue ethics (promoting the moral character), and the ethics of care, which emphasizes the empathy and the protection of vulnerable groups (Florridi, 2018). These frameworks contribute to the development of a moral community involving developers, users, and beneficiaries, negotiating social norms for the appropriate use of AI („Society-in-the-loop"), ensuring a relationship of trust based on moral values and social responsibility (Ferdaus, 2024).

A framework for developing a trustworthy AI system with a multidisciplinary approach

To ensure the development of a trustworthy AI system that integrates multidisciplinary expertise, the proposed framework builds upon the established standards (EU AI Act, ISO/IEC 42001, IEEE 7000, and NIST AI RMF 1.0) and extends them through six iterative steps:

Step 1. Defining purpose, accountability, and scope: Establish clear objectives and accountability mechanisms for AI development, assigning responsibilities to stakeholders. This stage requires collaboration among experts in technology, ethics, law, sociology, psychology, and public policy to ensure that the objectives and risk assessments reflect the societal values and diverse perspectives from the outset.

Step 2. Assessing the impact on fundamental rights: Conduct a multidisciplinary impact assessment to identify the potential effects on human rights, equality, and data protection.

The integration of ethical, legal, and social perspectives enables an early identification of risks such as bias, discrimination or privacy violations, especially for vulnerable groups.

Step 3. Ethical and inclusive design: Embed ethical principles and inclusiveness in AI system design through participatory engagement with the stakeholders across the disciplines and societal sectors. Ensure the transparency of the design choices, traceability of data sources, and alignment with pluralistic value systems to enhance legitimacy and fairness.

Step 4. Testing, validation, and verification: Evaluate the AI system through technical, ethical, and legal validation processes. Beyond the standard performance metrics, include adversarial robustness testing, ethical audits, and conformity assessments to ensure fairness, transparency, and compliance with the applicable standards and regulations.

Step 5. Controlled implementation and continuous monitoring: Deploy the system with active human oversight and multidisciplinary monitoring. Feedback loops involving technical, ethical, and legal experts ensure that the system adapts to the real-world conditions, identifies new risks, and maintains the compliance with the evolving norms.

Step 6. Audit, transparency, and communication: Implement regular interdisciplinary audits covering both the technical reliability and the ethical compliance. Maintain a transparent reporting to inform users, regulators, and the public. Clear communication fosters accountability, institutional trust, and informed governance.

This framework extends beyond the existing trustworthy AI models by operationalizing a multidisciplinary collaboration throughout the entire AI lifecycle. While many international guidelines, such as the EU AI Act, the OECD AI Principles, ISO/IEC 42001, and IEEE 7000, outline the general requirements for transparency, fairness, and accountability, they often treat the ethical, legal, and technical aspects in isolation. This framework introduces an integrative structure in which the experts from diverse fields (technical, ethical, legal, social, and psychological) participate in every phase, from design to auditing.

DIMENSIONS OF TRUST: A MULTIDISCIPLINARY APPROACH

Building trust in AI systems is a complex challenge that cannot be addressed only through technological components. It requires a broader perspective, integrating technical, ethical, social, and legal elements to form a coherent and sustainable framework for the responsible development and use of AI in the digital society. In this context, trust becomes a multidimensional concept, encompassing several key dimensions that need to be managed in an integrated manner to ensure that AI systems function not only correctly from a technical point of view, but also in accordance with people's fundamental values and rights.

The technical dimension of the trust starts from the need for AI systems to be robust and secure, for example, to operate reliably even in the face of unexpected conditions or disruptions, to withstand cyberattacks, and to avoid errors that could negatively affect users or society. The technical robustness implies the ability of AI to ensure the continuity of its operations without prejudice, so that the results are predictable and reliable. At the same time, the transparency and explainability of AI decisions become critical for users, auditors, and authorities who need to understand the reasoning behind the automated decisions. XAI reveals the internal mechanisms of AI systems, and allows an accurate assessment of the correctness and legality of the decision-making processes.

The ethical dimension of the trust is closely linked to the respect for fundamental rights and the moral principles that guide the interaction between people and technology. Equity and non-discrimination, as fundamental principles, require that AI systems treat all users fairly and impartially, without reproducing or emphasizing pre-existing biases and stereotypes in the training data or in the design of the algorithms. This dimension derives from

various ethical theories: the deontology, which emphasizes the respect for the person and moral obligations; the utilitarianism, which evaluates the consequences of actions in maximizing the common good; the virtue ethics, which emphasizes the formation of moral character; the social contract, which underpins consensual rules and norms; and the ethics of care, which emphasizes the responsibility towards vulnerable groups and the contextual relationships in the use of AI. Responsibility, another principle, involves assuming the effects of automated decisions, with clear mechanisms for redressing errors and abuses. The legal and governance dimension helps establish a stable, clear, and enforceable framework that imposes the compliance with norms, regulations, and fundamental rights at all stages of the AI lifecycle (from design, development, and testing, to deployment and monitoring). The emphasis on the legislative framework, such as AI Act, GDPR, and other European regulations, supports the concrete and verifiable application of the concept of trustworthy AI, providing a mechanism for accountability and external control. Last but not least, governance involves continuous human oversight, regular audits, and the promotion of genuine technical and institutional transparency (AI Act, 2024).

The social dimension reflects the acceptance and trust of users and society as a whole. This includes the involvement of different social groups and vulnerable communities in the development, validation, and evaluation of AI systems, ensuring that the technology responds to needs and respects the diversity of values. Education and transparent communication are important for building an informed and authentic trust, allowing users and decision-makers to understand the benefits and risks of using AI, as well as their rights in this context. To integrate these dimensions, a collaborative multidisciplinary approach is appropriate, involving experts from various fields: technical developers, ethicists, lawyers, sociologists, psychologists, public policy experts, and representatives of civil society. This collaboration ensures a holistic perspective that can respond to the complexity and interdependencies that characterize AI systems in the real world (Vincent-Lancrin & van der Vlies, 2020).

This integrative and multidisciplinary approach (based on main dimensions of trust - human supervision, technical robustness and safety, privacy and data governance, equity and non-discrimination, transparency and explainability, responsibility and auditability, social impact and well-being, inclusion and diversity) has built the foundation for the development of trustworthy AI, reflected in various international initiatives and standards, which aim to transform ethical principles into clear, enforceable and verifiable norms. Ultimately, trust in AI becomes a shared construct, grounded in collaboration, dialogue, and respect for human values, integrating technology in a sustainable and responsible way into the digital society.

STANDARDIZATION INITIATIVES AND AI GOVERNANCE

With a rapid evolution of AI and expansion of its applicability in the areas of life, standardization and governance initiatives have acquired special importance for the development and responsible use, safe and transparent use of AI technologies. Thus, a unified framework of rules and best practices became necessary to guarantee the interoperability, security, ethics, and respect for fundamental rights in the use of AI. AI governance integrates legal regulations, technical standards, and oversight mechanisms, capable to ensure the compliance and protection of the users.

AI standardization

Operationalizing the concept of trustworthy AI requires technical standardization by transforming ethical and legal principles and recommendations into measurable and consistently applicable requirements in the process of designing, developing, testing, and implementing AI systems. Several important bodies define the relevant standards; these include: the International Organization for

Standardization (ISO), the International Electrotechnical Commission (IEC), the Institute of Electrical and Electronics Engineers (IEEE), and the National Institute of Standards and Technology (NIST) (see ISO/IEC. 2020, 2021, 2022; IEEE Standards Association, 2021, 2022; NIST, 2023).

Basic standards and guidelines for trustworthy AI are the following, as can be seen in Table 4 (OECD, 2025; ETSI, 2024-2025):

Table 4. *Technical and ethical standards for AI Trustworthy*

Standard	Regulates	Main purpose	Practical utility	Specific scope	Risks
ISO/IEC 42001	Set of requirements for an integrated AI management system, including policies, processes, audit, and accountability	Governance and organizational control for AI	Provides a framework for organizations developing AI in accordance with the AI Act and ethical principles	AI-based services in the fields of finance, health, technology, and public services	Governance, organizational control, compliance assurance, accountability, and management of ethical issues, transparency, and respect for user rights
ISO/IEC TR 24028-24029-1	Framework for assessing AI trust and safety, providing guidelines to define, measure, and manage risks related to security, reliability, and transparency in AI systems	Enhancing the safety and reliability of AI; testing the robustness of AI models	Help to implement and monitor compliant and reliable AI	Neural networks in AI systems in critical infrastructures, measuring stress and ensuring AI continuity in operational environments, sensitive applications	Governance, accountability, compliance with regulations in force, ethics, and responsible operation
IEEE7000 - 7002	Ethical aspects, responsibility, transparency, and governance, respecting user rights, and avoiding bias and risks in applications	Guides ethically, transparently, securely, and autonomously the development of systems, respecting users' rights	Provide a clear and applicable framework for integrating ethical and trust principles into the development of AI systems	Systems and software engineering, preventing risks related to privacy violations and ensuring respect for user rights	Violation of user rights, lack of transparency in automated decisions, algorithmic bias, unclear responsibility in case of errors or abuses

NIST AI RMF 1.0	Identifying, assessing and managing AI risks (bias, robustness, human control)	Reducing risks associated with AI	Provides frameworks, standards, and technical guidelines for information security management	Credit scoring, fraud detection, AI-assisted medical diagnosis, predictions in medicine, industrial automation, security systems, and sensitive data management	Algorithmic discrimination, errors in automated decisions, lack of transparency, and cybersecurity vulnerabilities
OECD AI	Principles and good practices for AI development, implementation, and governance	Global ethical foundation	Promotes the development, implementation, and responsible governance of AI	Health, education, public services, finance, and digital technologies	Unethical use of AI, violation of fundamental rights, lack of transparency, lack of accountability, discrimination, and inequality in access to technology
ETSI EN 303 645	Basic requirements and good practices in the field of cybersecurity for IoT devices	Protecting security and data in IoT and AI	Establishes a uniform set of cybersecurity requirements for Internet-connected devices	Security and data protection of IoT devices connected to network infrastructures	Security and protection of user data, cyber vulnerabilities, intrusion attacks, compromising the confidentiality and integrity of information, unauthorized access to devices and networks

The table reflects a clear picture of the different types of standards that must be followed in building reliable AI systems. The specific areas of use and risks that may arise are highlighted. AI Act, together with technical standards, plays a supporting role in operationalizing these rules. The compliance with these standards contributes to improving the quality of AI systems.

AI governance

AI governance encompasses the system of policies, regulations, procedures, and institutional mechanisms that ensure a transparent and ethical use of AI. This involves continuous monitoring, external audits, human oversight, and the facilitation of the dialogue between

the actors involved in the development and application of AI systems: public sector, private sector, academia, and civil society.

In Europe, the governance is based on rigorous legal regulations, technical standardization norms, and voluntary initiatives that promote cooperation and exchange of good practices between various stakeholders, facilitating the efficient and compliant implementation of AI in the digital space.

In Romania, the efforts to align these regulations and international standards are supported by the activity of the National Standardization Body (ASRO), which operates through the National Technical Committee ASRO CT 401 dedicated to AI. It brings together experts from local industry, academia, and government, supporting the development of a competitive and internationally compliant AI ecosystem. At the corporate level, AI governance includes the stability of internal codes of conduct and best practices aimed at risk assessment, cybersecurity, transparency of automated decisions, and organizational readiness for human intervention. This ensures that AI technologies are responsibly integrated into decision-making and operational processes, in accordance with ethical principles and legal regulations.

Globally, the future governance is a main topic on the scientific forums and conferences agenda. Specialists are exploring innovative ways to conduct the governance process in a decentralized, transparent, and accountable manner, making use of emerging technologies such as AI and blockchain. Thus, the response to complex digital challenges is influenced in a substantial manner by following decentralized and participatory governance and the transformation of the state-citizen relationship. As a conclusion, one can say that current initiatives for standardization and governance reflect a growing global awareness for the need of a robust, harmonized, and inclusive framework for AI. This framework aims to ensure and use AI in a safe, ethical, and responsible manner, with fundamental rights protected, facilitating equity and social inclusion. A multidisciplinary and collaborative approach is largely capable of adequately responding to the diverse and complex impacts of AI in digital society. Table 5 presents the legal framework for AI, which includes technical and ethical standards.

Table 5. *Overview on criteria and standards - the legal framework for trustworthy AI*

Criteria	ISO / IEC	IEEE (7000)	NIST	AI Act
Type	International technical standard	Ethical and technical standard	Risk management framework	Binding legislative regulation
Objective	Secure and auditable AI governance	Integrating ethical values into AI	AI risk identification and mitigation	Uniform regulation for trusted AI
Applicability	Global (technical)	Global (technical-ethical)	US (and international)	EU area (with extraterritorial effect)
Areas covered	Organizational AI systems	Transparency, bias, confidentiality	AI Governance and Risk Assessment	High-risk AI, deepfake, audit
Mandatory	Voluntary (compliance based)	Voluntary (ethical guide)	Voluntary (management tool)	Mandatory in the EU

AI risk	Robustness and audit approach	Ethical approach	Risk classification and management	Classification: unacceptable, high-risk, etc.
Human oversight	Recommended	Promoted in design and use	Part of risk assessment	Mandatory for high-risk AI
Conformity assessment	Internal and external audit	No formal certification	No certification mechanism	Strong requirements and external control
Error reduction rate	AI error rate <1%, Decision transparency >95%, Incident response time <24h	Algorithmic bias <5%, ethical compliance 100%, user satisfaction >90%	Major risk reduction >80%, quarterly risk assessment, remediation in max 48h	Annual external audit of high-risk AI, transparency and accountability 100%, incident reporting <72h
Percentage of security incidents reduced	60-70%	40-60%	>75%	>70%
Degree of transparency in decisions	>90%	>85%	>90%	100%
Effectiveness of risk assessments / frequency of early identification and remediation of risks	>80%	30%	>80%	>70%

EXAMPLES OF TRUSTED AI APPLICATIONS IN VARIOUS FIELDS

The development and implementation of trustworthy AI systems is a key priority for the smooth and responsible integration of AI technologies into society. This trust is not just a technical requirement, but a social, ethical, and legal imperative, which needs to be demonstrated concretely through practical achievements in key sectors, from health to finance, education, public administration, and industry. In this section, relevant examples are presented to illustrate how the principles of trustworthy AI (robustness, transparency, fairness, accountability, data protection, and human oversight) are translated into operational realities in various critical areas.

Health

The healthcare sector is one of the most sensitive and dynamic areas where AI can bring major benefits, but also significant potential risks. Here, AI systems must deliver accurate and reliable results, protect confidential patient

data, and allow for human intervention when necessary.

A concrete example is Kepler Night Nurse AI, an advanced system for monitoring patients in hospitals and at home, which focuses on preventing critical nocturnal incidents, such as falls or loss of orientation. This system uses sophisticated computer vision and AI technologies to detect risky behaviors in real time, generating rapid alerts for medical staff, but with a low degree of false alarms, which reduces operator stress. The implementation of Kepler Night Nurse AI was carried out with increased attention to compliance with data protection legislation (GDPR) and security standards (ISO 27001), as well as through rigorous external audits to validate the algorithms (Zhang and Zhang, 2023).

In addition to patient safety, another advantage is the facilitation of continuous monitoring without affecting patient privacy and without generating unnecessary interventions, thus increasing the efficiency of healthcare. Its technical features include the possibility of operating in the cloud or on-premises (edge computing), robust scalability, and integration with other existing medical systems for an optimized workflow.

Detailed expansion and technical inserts – Kepler Night Nurse AI

General description:

Kepler Night Nurse AI is an intelligent video patient monitoring system developed to prevent accidents such as falls or disorientation of patients in hospitals and at home, especially at night. This system uses computer vision and AI algorithms to detect risky behavior in real time (Kepler).

Technical architecture:

- Hardware components: HD IP video cameras installed in monitoring areas, edge computing equipment that preprocesses data locally to reduce latency and data privacy risks;
- AI processing: use of specialized deep learning models for body position recognition and detection of behavioral anomalies (e.g., falls, unusual movements);
- Data flows and alerting: processed data generate automatic alerts that are integrated into existing nurse call systems. The generated reports are transmitted in real time to the medical staff, with the option of aggregation for further analysis;
- Infrastructure: the system can operate in hybrid mode, using both edge computing for local processing of sensitive data and cloud for scaling, storage, and additional analysis, complying with GDPR requirements and data protection regulations.

Applicable standards:

- ISO/IEC 27001 for information security management, guaranteeing the protection of medical data;
- GDPR for the protection of personal data;
- ISO/IEC TR 24028:2020 on the robustness and transparency of AI systems, applicable in the algorithmic component to ensure the reliability of the system;
- Periodic external audits according to trust criteria (robustness, transparency, accountability).

AI trust aspects implemented:

- technical robustness for correct detections even in low light conditions or atypical positions.
- transparency by explaining the alerting criteria used;
- human supervision with the possibility of medical personnel to intervene or deactivate the system;
- data protection and confidentiality through local processing and encryption of video streams.

Finance

In the financial sector, the challenges related to fairness, transparency, and security are particularly acute, given the high degree of regulation and the importance of the impact of financial decisions on the lives of individuals and companies.

Supply chain finance and receivables management platforms exemplify how

trustworthy AI can transform financial processes. These systems automate the entire lending and collection cycle, monitor risks in real time, and integrate advanced predictive analytics and decision support functions. Transparency mechanisms provide users with clear explanations of automated decisions regarding the approval or rejection of a loan, and the permanent auditability ensures continuous monitoring of the algorithms' compliance with legal regulations.

Furthermore, implementing a human intervention system to review the automated decisions is important to prevent errors and guarantee fair treatment, especially when targeting vulnerable or high-risk groups. These solutions contribute to reducing discrimination, one of the most critical risks in the financial sector, ensuring a broader financial inclusion and a more stable economic environment.

Detailed expansion and technical inserts – Qualco

General description:

In the financial sector, AI solutions are used to automate credit and receivables management processes, predictive analytics for risk management, and financial flow optimization. Qualco platform integrates automated decision engines, omnichannel management, and legal facilities adapted to the specifics of the market.

Technical architecture:

- Data and integration: collection of internal and external financial data (financial reports, behavioral data, bank scoring), managed in secure big data platforms;
- AI processing: machine learning ML models for predictive risk analysis, advanced segmentation, and optimization of lending decisions;
- Automation: automated workflows, including notifications, status updates, and report generation, all integrated through APIs with existing financial and legal systems;
- Human intervention: Systems incorporate modules for automatic review of decisions by human experts, to ensure control and fairness.

Applicable standards:

- ISO/IEC 42001:2023 for AI system management, covering all aspects of governance, audit, and accountability;
- NIST AI risk management framework (2023) for identifying and managing the risks related to the distribution of automated decisions;
- EU regulations, including the AI Act on transparency, audit, and risk classification;
- ISO/IEC TR 24029-1:2021 for ensuring the robustness of the models used.

AI trust aspects implemented:

- transparency of AI-based decisions, with clear explanation of the reasons for credit approval/denial;
- auditability and full logging of all automated decisions;
- protection of personal data and confidentiality of transactions, in accordance with GDPR;
- human oversight in critical decisions, to prevent discrimination and errors.

Education and Human Resources

Applying the trustworthy AI to education and human resources addresses the need for fair, transparent, and accountable talent assessment, recruitment, and management processes. In these areas, decisions can affect not only careers or educational opportunities, but also social equity and inclusion.

The HireVue platform illustrates an advanced AI model used for evaluating job candidates, combining scoring algorithms with decision explainability. The system provides clear and transparent criteria for scoring candidates, and the possibility of human intervention (HITL - Human-in-the-Loop) allows for review of decisions, especially in situations of ambiguity or uncertainty. Periodic external audits verify and reduce algorithmic bias related to gender, race or nationality, aligning it with GDPR and EEOC standards. Also interesting is the platform's participatory approach, which involves multiple stakeholders, including HR specialists, psychologists, and representatives

of vulnerable groups, in the process of defining and updating the criteria and values used in a candidate evaluation. The modular and scalable technological architecture ensures the adaptability and easy integration with existing workforce management systems.

Detailed expansion and technical inserts – HireVue

General description:

HireVue is an advanced AI system used to select and evaluate job candidates, using an automated audio-video analysis and cognitive tasks, with an increased focus on transparency and fairness.

Technical Architecture:

- Data processing: The system collects and analyzes audio-video responses of the candidates, using neural networks for natural language processing, facial expression recognition, and voice tone analysis;
- Score modeling: Scoring algorithms use multiple factors to create objective profiles of the candidates, with clear and explainable criteria;
- Cloud-based platform: Allows scalability and integration with existing HR systems (ATS – Applicant Tracking Systems);
- HITL: Automated assessments can be fully or partially reviewed by human assessors, and the system prioritizes the cases that require additional attention.

Applicable standards:

- IEEE 7000-2021 Ethically Aligned Design on integrating ethics into the AI lifecycle, especially fairness and non-discrimination;
- GDPR and EEOC for data protection and compliance with anti-discrimination rules;
- IEEE 7001-2021 Transparency of Autonomous Systems for ensuring the explainability of automated decisions;
- Periodic audits to assess and reduce bias, including model reviews and testing on diverse groups.

AI trust aspects implemented:

- high explainability of decisions, with detailed feedback for candidates and recruiters;
- human control involved in key decisions through the HITL mode;
- protection of sensitive video and audio data, encryption, and secure storage;
- involvement of the society and relevant stakeholders in the evaluation and adjustment of the criteria.

Public administration and justice

The public sector, especially the areas of administration and justice, is a delicate terrain for implementing AI, due to the direct impact on citizens' rights and freedoms. AI systems used here must rely on the highest degree of transparency, accountability, and human control.

In Romania, the use of algorithms in the allocation of resources such as social funds, aid programs, and school admissions has highlighted significant problems related to the lack of explainability of automated decisions and the difficulty of human control. In many cases, such systems have generated perceptions or effects of discrimination, and the mechanisms for contestation or auditing are currently underdeveloped.

To restore trust, it is essential to integrate trustworthy AI principles such as the possibility of challenging automated decisions, independent external audits, and continuous human oversight. Also, the compliance with the European legislative framework (AI Act, GDPR) provides clear and rigorous support for the development of an AI ecosystem that serves the public interest in a fair and responsible manner.

Detailed expansion and technical inserts – Public Administration and Justice

Context and Challenges:

The use of AI in public administration and justice requires special attention, given the direct legal and social impact on citizens. The automated initiatives have targeted the distribution of social funds, access to various public services, and admission to the educational system. The major problems consist

of the lack of transparency of the decisions or the difficulties of human supervision.

Proposed architecture and good practices:

- hybrid AI-human systems: automated decisions are supported by oversight teams for impact analysis and decision correction;
- open and auditable platforms: all decisions are documented and available for external audit, with real-time traceability;
- appeal mechanisms: citizens have the opportunity to request a review of automated decisions, and the process is managed transparently;
- compliance with legislation: integration of the AI Act and GDPR, with a fundamental rights impact assessment.

Applicable standards:

- AI Act - obligation of transparency and mandatory human oversight;
- ISO/IEC 42001:2023 for AI governance management in the public sector;
- OECD recommendations on AI ethics in the public sector to ensure fairness.

Industry and Cybersecurity

In modern industry and cybersecurity, AI plays a central role in production optimization, predictive maintenance, quality control, and protection of critical digital infrastructures. In this context, trust is closely linked to the technical robustness, transparency, and ability of AI systems to react correctly to cyberattacks or errors.

ETSI standards on security for IoT devices using AI provide a technical framework that directly contributes to ensuring security and data protection in these complex environments. The compliance with such standards and the AI Act requirements contributes to the development of resilient, audited, and guaranteed industrial solutions.

Specific challenges include continuous adaptation to new and sophisticated attacks, the clarity of automated decision-making processes used to detect and neutralize cyber threats, and facilitating an effective human intervention in critical processes.

Detailed expansion and technical inserts – Industry and Cyber Security

Practical application:

In industry, AI helps monitor and optimize production processes, quality assurance, and predictive maintenance. In cybersecurity, AI detects anomalies in network traffic, attacks, and vulnerabilities.

Architecture and standards:

- distributed AI/edge systems: local data processing for speed and privacy, with centralized control for management and auditing;
- continuous learning protocol: AI adapts in real time to new types of attacks or failures, through incremental learning models.

Standardization and compliance:

- ETSI EN 303 645 for security of IoT devices integrated with AI;
- ISO/IEC TR 24028 and NIST AI RMF for robustness and risk management;
- Implementation of AI Act requirements for oversight and auditing.

Trust and resilience:

- full auditability of AI decisions;
- human oversight with the possibility of rapid intervention;
- compliance with cybersecurity and data protection rules.

AI systems like Kepler Night Nurse AI and IBM Watson in health, HireVue in education, Qualco in finance, TRUST XAI in cybersecurity, COMPAS, HART in legal / justice, Siemens MindSphere, GE Predix in industry, illustrate the potential and challenges of trustworthy AI, especially regarding explainability, fairness, and measurable impact.

Some of the strengths of trustworthy AI of these tools are: explainability and transparency, domain-specific impact, and measurable outcomes. XAI methods are increasingly adopted in healthcare, finance, industry, and justice to make AI decisions more interpretable for users. For example, in healthcare (Kepler Night Nurse AI, IBM Watson), XAI helps clinicians understand AI-driven diagnoses, supporting

better decision-making and error identification (Hossain et al., 2023; Borys et al., 2023; Albahri et al., 2023; Kalasampath et al., 2025).

In finance (Qualco), education (HireVue), and industry (Siemens MindSphere, GE Predix), XAI techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide actionable insights, improving stakeholder confidence and regulatory compliance (Kalasampath et al., 2025). These techniques make AI models more transparent and understandable to humans.

Trustworthy AI systems are evaluated using metrics such as accuracy, robustness, bias risk, and user satisfaction. In healthcare, measurable improvements include increased diagnostic accuracy and reduced health risks; in industry, enhanced operational efficiency and predictive maintenance are key outcomes (Afroogh et al., 2024; Albahri et al., 2023; Kalasampath et al., 2025).

Some of the weaknesses and challenges of trustworthy AI of these tools are: black-box complexity, bias and fairness risks, limited standardization, and user acceptance. Many AI systems remain opaque; thus, it is difficult for users to understand or challenge decisions, especially in high-stakes domains like justice and healthcare. This opacity can undermine trust and hinder adoption (Hossain et al., 2023; Afroogh et al., 2024; Borys et al., 2023; Kalasampath et al., 2025).

AI models in legal / justice (COMPAS, HART) and finance (Qualco) are susceptible to bias, potentially leading to unfair or discriminatory outcomes. High bias risk and low methodological quality are persistent concerns, particularly in healthcare applications (Afroogh et al., 2024; Albahri et al., 2023; Kalasampath et al., 2025). There is a lack of standardized, domain-specific metrics for evaluating trustworthiness, making cross-sector comparisons challenging. Many systems lack rigorous, transparent evaluation frameworks (Afroogh et al., 2024; Albahri et al., 2023).

The trust in AI is influenced by explainability, perceived fairness, and cultural factors. Distrust can arise from perceived threats to autonomy, privacy or dignity, especially when AI systems are used for surveillance or decision-making without adequate human oversight (Afroogh et al., 2024; Rădulescu et al., 2025). In Table 6, a comparison of trustworthy AI strengths, weaknesses, and outcomes is presented.

Table 6. *A comparison of trustworthy AI strengths, weaknesses, and outcomes*

Sector/AI System	Strengths (Trustworthy AI)	Weaknesses (Trustworthy AI)	Measurable / Comparative Outcomes	Citations
Healthcare (Kepler Night Nurse AI, IBM)	XAI improves diagnosis, transparency	Black-box risk, bias, low quality	Diagnostic accuracy, bias risk, user trust	(Hossain et al., 2023; Borys et al., 2023; Albahri et al., 2023; Kalasampath et al., 2025)
Finance (Qualco)	Actionable insights, compliance	Bias, lack of standard metrics	Fairness, regulatory adherence	(Afroogh et al., 2024; Kalasampath et al., 2025)
Education (HireVue)	Automated assessment, transparency	Bias, explainability	Fairness, user satisfaction	(Afroogh et al., 2024; Kalasampath et al., 2025)
Justice (COMPAS, HART)	Decision support, efficiency	Opaqueness, bias, fairness issues	Bias audits, fairness metrics	(Afroogh et al., 2024; Kalasampath et al., 2025)

Industry (Siemens, GE)	Predictive maintenance, efficiency	Explainability gaps	Operational efficiency, error reduction	(Kalasampath et al., 2025)
Cybersecurity (TRUST XAI)	Transparency, risk assessment	Complexity, explainability limits	Threat detection accuracy, trust scores	(Afroogh et al., 2024; Kalasampath et al., 2025)

The examples presented clearly show that trustworthy AI is not a simple technical or legal aspiration, but a multidimensional reality that requires an integrated approach. Successful application of the principles of trustworthy AI is possible only through an interdisciplinary collaboration between developers, ethicists, lawyers, users, and policymakers.

Responsible implementation of AI in key areas ensures not only technological performance, but also social acceptability, respect for fundamental rights, and promotion of inclusion. This becomes an essential condition for the ethical and sustainable integration of AI in contemporary society.

The detailed examples illustrate how the principles of trustworthy AI are embodied in robust technological solutions, scalable and interoperable architectures, high-performance audit and transparency mechanisms, effective data protection measures, and the constant application of human oversight. Their correct implementation contributes to increasing user and societal trust in AI, facilitating large-scale adoption and reducing the risks of abuse or errors.

CHALLENGES AND FUTURE DIRECTIONS IN DEVELOPING TRUSTWORTHY AI

As AI continues to advance rapidly and become more deeply integrated into all aspects of economic, social, and cultural life, new directions and major challenges in ensuring trustworthy AI are emerging. The future of sustainable development of AI technologies must consider these challenges to guarantee safety, fairness, transparency, and respect for fundamental rights in all AI applications.

Major challenges

The algorithmic bias and the inequities are the most pressing issues. Training data can contain historical or cultural prejudices that can be reflected in automated decisions, generating discrimination that is difficult to detect and combat. In addition to the nature of the data, the concepts and models on which the algorithms are based can implicitly incorporate this bias, disproportionately affecting certain social groups, including the vulnerable ones. Thus, managing bias is not only a technical issue, but also an ethical and social one, requiring continuous audits, diversification of the data used, assessments on the marginalized groups, and the development of robust correction methods.

Data underrepresentation is an extension of the problem of bias, whereby certain segments of the population or particular situations are poorly represented in the data sets used to train AI models. This can lead to frequent errors, inappropriate decisions, and marginalization. Correcting this problem requires dedicated efforts to collect and integrate representative data, including through collaborations between research institutions, the public sector, and the targeted communities.

The lack of transparency and explainability in many AI systems, especially those based on complex techniques such as deep neural networks, remains a significant barrier to trust and accountability. Users, implementers, and regulators need clear, accessible, and understandable explanations of how systems make decisions so they can assess their correctness, legality, and ethics. Further development of XAI techniques is required in order to overcome these limitations.

Insufficient human control and oversight are vulnerabilities that can generate major risks, including inappropriate automated decision-making, abuse or unidentified errors. Effective mechanisms for involving the human factor, both in the decision-making process and in post-implementation monitoring, are vital to maintain safety and accountability of AI use, especially in critical applications.

The lack of a globally harmonized regulatory framework creates difficulties in standardizing requirements for trustworthy AI. The differences between national or regional regulations, such as those between the EU, the USA or China, can create interoperability issues, hinder innovation, and reduce user protection. International collaboration and the adoption of common principles and standards are imperative to overcome these obstacles.

The challenges related to generative AI and general-purpose AI are increasingly significant, given their potential to generate false or manipulative content (deepfake, fake news), to affect democratic processes and public trust, as well as to produce complex negative effects that are difficult to anticipate. Addressing these risks requires the development of detection mechanisms, restrictions on use, responsible policies, and the active involvement of civil society.

The socio-economic and workforce impact is a major unknown. Accelerated automation and the adaptation of societies to new digital models can create inequalities, technological unemployment, and the need for extensive professional retraining and digital inclusion programs. Building trustworthy AI must also integrate this dimension, promoting just social transitions.

Directions and proposed solutions

To respond to the increasing complexity, research and development of trustworthy AI is focused on:

- multidisciplinary and collaborative approaches that integrate skills from technology, ethics, law, sociology, and psychology, for a holistic and applicable understanding of AI challenges. Projects should involve not only engineers, but also experts in social and humanities fields, as well as representatives of civil society;
- continuous development of audit and assessment techniques for protection against bias, algorithmic transparency, and robustness of AI systems. Tests and validations should be permanent, with the involvement of independent third parties to guarantee impartiality;
- expanding and strengthening the regulatory framework and global standardization. Initiatives such as the AI Act in the EU are becoming reference models, along with ISO, IEEE, and NIST standards, which need to adapt increasingly quickly and be supported by real implementation and control mechanisms;
- promoting multidisciplinary education and digital literacy for developers, users, policymakers, and the general public. Only through knowledge and understanding of the functioning of AI, the risks and rights, can a culture of trust and responsibility be built;
- implementing robust human oversight mechanisms to ensure effective intervention and constant checks of automated decisions, especially in sensitive areas;
- responsible management of social impact and economic changes, through retraining, inclusion, and support policies for vulnerable groups affected by accelerated digital transformations;
- developing technical tools for explainable and transparent AI, to be natively integrated into design and operational processes, facilitating access to clear information for all actors involved.

The importance of continuous social dialogue and civic engagement

An essential aspect of building trustworthy AI is openness to dialogue between the developers, institutions, academia, non-governmental organizations, and citizens.

A comprehensive understanding of the societal concerns, combined with the inclusion of diverse values into the design and regulation of AI, is vital for the acceptability and sustainability of this technology. Actively engaging vulnerable groups and diverse stakeholders helps prevent exclusion and discrimination, ensuring that AI technologies authentically serve the interests of the communities and society as a whole.

Research and innovation perspectives

The future of trustworthy AI is closely linked to the progress in areas such as:

- XAI, which provides advanced methods for interpreting and making automated decisions transparent;
- Robust and resilient systems capable of operating reliably in complex environments and resisting sophisticated cyberattacks;
- Automated and semi-automated mechanisms for auditing and managing algorithmic risks;
- Ethical technologies integrated at the design level, including continuous monitoring of social effects and accessible mediation of AI performance.

These directions ensure not only the technical performance of AI systems, but also their ability to adapt to new ethical, social, and legal challenges. The development of trustworthy AI in the context of digital society is a continuous, multidimensional, and collaborative process, which requires the integration of technical, ethical, social, and legislative perspectives. In the face of emerging challenges (algorithmic bias, lack of transparency, social, and economic risks), only a responsible approach, applied through unitary standards, education, and permanent dialogue with society, can ensure that AI becomes a beneficial, equitable, and trustworthy tool for citizens.

CONCLUSIONS

The paper provides an analysis of the concept of trustworthy AI from a multidisciplinary perspective, highlighting the importance of integrating technical, ethical, legal, and social dimensions in the development, implementation, and governance of AI systems. Against the backdrop of the widespread adoption of AI technologies in different critical areas (health, finance, education, public administration, and industry), ensuring a high level of trust becomes a condition for their acceptable and responsible use.

The development of trustworthy AI faces persistent challenges, such as algorithmic bias, data underrepresentation, lack of explainability and difficulties of human oversight, as well as the complexity of the global legal framework. These require multidisciplinary, collaborative approaches, including a transdisciplinary education, the continuous development of explainable and robust techniques, harmonized policies, and an open social dialogue that recognizes and integrates the diversity of values and needs of societies.

The unique aspects that the paper brings are: a multidisciplinary synthesis, a coherent framework, and relevance to the European / Romanian context. Instead of focusing on trustworthy AI from a single discipline (e.g., purely technical transparency or purely legal compliance), the work integrates contributions from five distinct fields: computer science, ethics, law, psychology, and public policy. The paper connects trustworthy AI principles (like fairness and accountability) with concrete, auditable practices. This moves beyond listing principles by outlining how transparency is tied to explainability (technical requirement) with accountability that is supported by auditability (governance requirement). The proposed framework is designed to support responsible governance for building and evaluating trust across the entire AI lifecycle.

The paper grounds the discussion by analyzing the challenges of trustworthy AI adoption within the context of specific EU-based emerging digital economies. By referencing the activities of major companies in the Romanian digital sector (e.g., BCR, ING), the paper situates the analysis within an economy integrating AI under the regulatory pressures of the EU AI Act.

However, the paper does not address in depth the aspects related to generative AI and its emerging risks. Negative or critical case studies

that highlight failures or application limitations of current initiatives are not addressed. Not enough regional or cultural approaches are included in the multidisciplinary analysis of AI trust. Future research can be expanded in these areas to increase the applicability of the analysis provided by the paper.

ACKNOWLEDGEMENTS

This work was carried out through the Core Program within the National Research Development and Innovation Plan 2022-2027, carried out with the support of MCID, project no. 23380101, „Contributions to the consolidation of emerging technologies specific to the Internet of Things and complex systems".

REFERENCE LIST

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024) Trust in AI: Progress, challenges, and future directions. H*umanities and Social Sciences Communications*. 11(1), 1–30. doi: org/10.1057/s41599-024-03183-4

AI Act. (2024) *EU AI Act whitepaper*. www.age.bsigroup.com/eu-ai-act/whitepaper [Accessed 2nd July 2025]

Ahmed, M., Begum, S., Barua, S., Masud, A., Di Flumeri, G., & Navarin, N. (2025) Enhancing Explainability, Robustness, and Autonomy: A Comprehensive Approach in Trustworthy AI. In *Proceedings of the 2025 IEEE Symposium on Trustworthy, Explainable and Responsible Computational Intelligence (CITREx)* (pp. 1–7). doi: org/10.1109/CITREx64975.2025.10974944

Albahri, A., Duhaim, A., Fadhel, M., Alnoor, A., Baqer, N., Alzubaidi, L., Albahri, O., Alamoodi, A., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y., & Deveci, M. (2023) A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*. doi: org/10.1016/j.inffus.2023.03.008

Al-Kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024) Ethical challenges and solutions of generative AI: An interdisciplinary perspective. *Informatics,* 11(3), 58. doi: org/10.3390/informatics11030058

Ali, R., Hussain, A., Nazir, S., Khan, S., & Khan, H. (2023). Intelligent decision support systems—An analysis of machine learning and multicriteria decision-making methods. *Applied Sciences*. 13(22), 12426. doi: org/10.3390/app132212426

Alsalem, M., Alamoodi, A., Albahri, O., Albahri, A., Martínez, L., Yera, R., Duhaim, A., & Sharaf, I. (2024) Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach. *Expert Systems with Applications*. 246, 123066. doi: org/10.1016/j.eswa.2023.123066

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020) Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*. 20(1), 310. doi: org/10.1186/s12911-020-01332-6

Bareis, J. (2024) The trustification of AI. Disclosing the bridging pillars that tie trust and AI together. *Big Data & Society* 11, 1–15. doi: org/10.1177/20539517241249430

Borys, K., Schmitt, Y., Nauta, M., Seifert, C., Krämer, N., Friedrich, C., & Nensa, F. (2023) Explainable AI in medical imaging: An overview for clinical practitioners - Saliency-based XAI approaches. *European Journal of Radiology*. 162, 110787. doi.org/10.1016/j.ejrad.2023.110787

Chamola, V., Hassija, V., Sulthana, A., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*. 11, 78994–79015. doi: org/10.1109/ACCESS.2023.3294569

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M. L., Herrera-Viedma, E., & Herrera, F. (2023) Connecting the dots in trustworthy AI: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*. 99. 101896. doi: org/10.1016/j.inffus.2023.101896

Digital Strategy. (n.d.) https://digital-strategy.ec.europa.eu/ [Accessed 9th July 2025]

ETSI TR 104 031 V1.1.1. (2024-02) *Securing Artificial Intelligence; Collaborative Artificial Intelligence*. https://www.etsi.org/deliver/etsi_tr/104000104099/104031/01.01.01_60/tr_ 104 031v010101p.pdf/ [Accessed 9th July 2025]

ETSI TR 104 032 V1.1.1. (2024-02) *Securing Artificial Intelligence; Traceability of AI Models.* https://www.etsi.org/deliver/etsi_tr/104000_104099/104032/01.01.0160/tr_104032v010101p.pdf. / [Accessed 9th July 2025]

ETSI TR 104 066 V1.1.1. (2024-07) *Securing Artificial Intelligence; Security Testing of AI.* https://www.etsi.org/deliver/etsi_tr/104000_104099/104066/01.01.0160/tr_104066v010101p.pdf/ [Accessed 9th July 2025]

ETSI TR 104 051 V1.1.1. (2025-06) *Securing Artificial Intelligence; Security aspects of using AI/ML techniques in telecom sector.* https://www.etsi.org/deliver/etsi_tr/104000104099/1040 51 /01.01.01_60/tr_1040 51v010101p.pdf/ [Accessed 9th July 2025]

Ferdaus, M., Abdelguerfi, M., Ioup, E., Niles, K. N., Pathak, K., & Sloan, S. (2024) Towards trustworthy AI: A review of ethical and robust large language models. *IEEE. arXiv.* doi: org/10.48550/arXiv.2407.13934

Finzel, B. (2025). Toward trustworthy AI with integrative explainable AI frameworks. *it – Information Technology.* 67, 20–45. doi: org/10.1515/itit-2025-0007

Florridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds & Machines*. 28, 689–707. doi: org/10.1007/s11023-018-9482-5

Herzog, C., Blank, S., & Stehl, B. C. (2025) Towards trustworthy medical AI ecosystems – A proposal for supporting responsible innovation practices in AI based medical innovation. *AI & Society*. 40, 2119–2139. doi: org/10.1007/s00146-024-02082-z

HireVue. (n.d.) https://www.hirevue.com/ [Accessed 3rd July 2025]

Hossain, M., Zamzmi, G., Mouton, P., Salekin, M., Sun, Y., & Goldgof, D. (2023) Explainable AI for medical data: Current methods, limitations, and future directions. *ACM Computing Surveys.* 57, 1–46. doi: org/10.1145/3637487

IEEE Standards Association. (2021) IEEE 7001-2021 – IEEE Standard for Transparency of Autonomous Systems. *Institute of Electrical and Electronics Engineers.* https://standards.ieee.org/ieee/7001/10310/

IEEE Standards Association. (2022) IEEE 7002-2022 – IEEE Standard for Data Privacy Process. *Institute of Electrical and Electronics Engineers.* https://standards.ieee.org/ieee/7002/10311/

ISO/IEC. (2020). ISO/IEC TR 24028:2020 — Information technology — Artificial intelligence — Overview of trustworthiness in AI. *International Organization for Standardization.* https://www.iso.org/ standard/77607.html

ISO/IEC. (2021) ISO/IEC TR 24029-1:2021 — AI — Assessment of the robustness of neural networks — Part 1: Overview. *International Organization for Standardization.* https://www.iso.org/standard/ 77608.html

ISO/IEC. (2023) ISO/IEC 42001:2023 – Artificial intelligence — Management system. I*nternational Organization for Standardization.* https://www.iso.org/standard/81230.html

Kalasampath, K., Spoorthi, K., Sajeev, S., Kuppa, S., Ajay, K., & Maruthamuthu, A. (2025) A literature review on applications of explainable artificial intelligence (XAI). *IEEE Access*. 13, 41111–41140. https://doi.org/10.1109/ACCESS.2025.3546681

Kaur, D., Uslu, S., Rittichier, K., & Durresi, A. (2022) Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR).* 55, 1–38. doi: org/10.1145/3491209

Kepler. Night Nurse. Kepler Vision. (n.d.) https://keplervision.eu/en/night-nurse/ [Accessed 3rd July 2025]

Kusters, R., Misevic, D., Berry, H., Cully, A., Cunff, L., Dandoy, L., Díaz-Rodríguez, N., Ficher, M., Grizou, J., Othmani, A., Palpanas, T., Komorowski, M., Loiseau, P., Frier, C., Nanini, S., Quercia, D., Sebag, M., Fogelman, F., Taleb, S., Tupikina, L., Sahu, V., Vie, J., & Wehbi, F. (2020) Interdisciplinary research in artificial intelligence: Challenges and opportunities. *Frontiers in Big Data*. 3. doi: org/10.3389/fdata.2020.577974

Kyriakou, K., & Otterbacher, J. (2023) In humans, we trust: Multidisciplinary perspectives on the requirements for human oversight in algorithmic processes. *Discover Artificial Intelligence.* 3(1), 44. https://doi.org/10.1007/s44163-023-00092-2

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2021). Trustworthy AI: From Principles to Practices. ACM Computing Surveys, 55, 1–46. doi: org/10.1145/3555803

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Jain, A., & Tang, J. (2021). Trustworthy AI: A computational perspective. *ACM Transactions on Intelligent Systems and Technology.* 14, 1–59. doi: org/10.1145/3546872

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024) Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology.* 15. doi: org/10.3389/fpsyg.2024.1382693

Malamuthu, B., Balakrishnan, T., Deepika, J., P, N., Venkataramanaiah, B., & Malathy, V. (2025) Explainable AI for decision-making: A hybrid approach to trustworthy computing. International Journal of Computational and Experimental Science and Engineering. doi: org/10.22399/ijcesen

Mitan, E. (2022) A Training Framework for Cybersecurity. *In EDULEARN22 Proceedings.* (pp. 9309-9315). IATED. doi: org/10.21125/edulearn.2022.2243

NIST. (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). *National Institute of Standards and Technology.* https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf [Accessed 11th July 2025].

OECD AI Principles, https://www.oecd.org/en/topics/sub-issues/ai-principles.html; [Accessed 10th July 2025].

Polemi, N., Praça, I., Kioskli, K., & Bécue, A. (2024) Challenges and efforts in managing AI trustworthiness risks: a state of knowledge. *Frontiers in Big Data*. 7. doi: org/10.3389/fdata.2024.1381163

Qualco. (n.d.) https://www.qualco.eu/ [Accessed 1st July 2025].

Rădulescu, C. Z., Rădulescu, M., Vevera, A. V., & Boncea, R. (2025) A hybrid approach based on the Two Weight Vectors and Extended TOPSIS methods with application in cybersecurity. *International Journal of Information Technology & Decision Making.* (in press) doi: org/10.1142/S0219622025501135

Rădulescu, C. Z., & Rădulescu, M. (2024) A hybrid group multi-criteria approach based on SAW, TOPSIS, VIKOR, and COPRAS methods for complex IoT selection problems. *Electronics.* 13(4), 789. https://doi.org/10.3390/electronics13040789

Rodríguez, N., Ser, J., Coeckelbergh, M., De Prado, M., Herrera-Viedma, E., & Herrera, F. (2023) Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion.* 99, 101896. doi: org/10.48550/arXiv.2305.02231

Sharma, S. (2024) Benefits or concerns of AI: A multistakeholder responsibility. *Futures*. doi: org/10.1016/j.futures.2024.103328

The Joint Research Centre: EU Science Hub. (n.d.) https://joint-research-centre.ec.europa.eu/scien tific- portfolios/ai-and-data_en [Accessed th July 2025]

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. Electronic Markets, 31, 447–464. https://doi.org/10.1007/s12525-020-00441-4

Vincent-Lancrin, S., & van der Vlies, R. (2020) Trustworthy artificial intelligence (AI) in education: Promises and challenges. *OECD Education Working Paper*. 218. OECD Publishing. doi: org/10.1787/a6c90fa9-en

Voronin, A., & Savchenko, A. (2024). Artificial intelligence in management problems. *International Scientific Technical Journal "Problems of Control and Informatics."* doi: org/10.34229/1028-0979-2024-3-6

Zhang, J., & Zhang, Z. M. (2023) Ethics and governance of trustworthy medical artificial intelligence. *BMC Medical Informatics and Decision Making*. 23, 7. doi: org/10.1186/s12911-023-02103-9

This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.