Exploring trustworthy artificial intelligence and its stakeholders: A literature review

Hussain, Jabbar[1,*,†], Koutsikouri, Dina[1] and Ljungberg, Jan[1]

[1]*Department of Applied IT, division Informatics, University of Gothenburg, Forskningsgången 6, 41756, Göteborg, Sweden*

Abstract

With the rapid advancement of artificial intelligence, the concept of trustworthy AI (TAI) has gained significant prominence, and various guidelines have emerged to direct the development of trustworthy systems. While research has produced valuable insights about TAI, we lack a comprehensive understanding of its characteristics and for whom the current TAI frameworks are important. In this paper, we address this challenge through a scoping review of the concept of TAI, proposing distinct characteristics of TAI based on the their themes and identifying the stakeholders for whom current frameworks are most relevant. This paper contributes to the literature on AI-systems development and deployment by offering a comprehensive understanding of trustworthy AI. In addition, it highlights the challenges of translating the concept of trustworthy AI into practice and what this means across different stakeholder groups.

Keywords

Trustworthy AI, Responsible AI, Explainable AI, Ethical AI, Stakeholders

1. Introduction

Artificial Intelligence (AI) has rapidly evolved into a transformative research field, attracting significant attention from both academia and industry. This growing interest is reflected in the expansion of the AI market, driven by advancements in machine learning, increasing availability of data, and a rising demand for intelligent systems across various industries [1, 2], including public sector organizations [3]. However, the successful integration of AI systems into organizations and society depends on the capacity to develop trustworthy AI (TAI), that is, systems that are transparent, accountable, and aligned with ethical principles [4]. Users must also trust these systems. Importantly, the level of trust that end-users have in an AI-system impacts the degree of adoption of the system [5]. Building TAI is a complex, multidimensional challenge. Trust in technology is only about how well we understand how a system works; it also depends on people's biases and attitudes towards technology. From an academic standpoint, ensuring TAI requires an interdisciplinary approach spanning computer science, social sciences, and ethics [6]. As AI technologies become increasingly embedded in everyday operations, the demand for TAI has surged. This evolution has made it more important to understand the pillars

SAIS2025: Swedish AI Society Workshop 2025, 16-17 June 2025, Halmstad, Sweden.

*Corresponding author.

† First author.

jabbar.hussain@ait.gu.se (H. Jabbar); dina.koutsikouri@ait.gu.se (K. Dina); jan.ljungberg@ait.gu.se (L. Jan)

0000-0003-1170-9069 (H. Jabbar); 0000-0002-9258-8117 (K. Dina); 0000-0003-0616-122X (L. Jan)

and characteristics of TAI, and hence how to cultivate it through development and use.

In recent years, numerous frameworks, guidelines, and reports have emerged to establish trust in AI. Various organizations worldwide have developed ethical guidelines and principles to establish trust in AI systems [7]. These include the EC ethics guidelines for trustworthy AI [8], the OECD principles on AI [9], the white house AI principles (U.S.) [10], the Chinese AI principles [11], the NIST's trust assessment methods [12], the UNESCO's recommendation on ethical AI (international) [13], the ISO standards for ensuring trust in AI [14], and DARPA's guidelines for explainable artificial intelligence (XAI) [15], among others. Despite this growing body of work, research on TAI remains fragmented and often lacks clear guidance on what TAI entails and for whom it is most relevant. This lack of coherence complicates the implementation of existing frameworks and guidelines, where AI development and adoption are widely studied. Addressing these gaps is essential for improving our understanding of TAI and refining its practical applications for key stakeholders, including users, developers, policymakers, and managers.

Although numerous guidelines exist, the conceptualization of TAI remains inconsistent across disciplines, and terminological variations further contribute to fragmentation. Even within the same terminology, different sources offer slightly different interpretations of TAI principles, requirements, and recommendations. This inconsistency fosters the perception that AI communities lack a unified understanding of TAI, complicating efforts to build trust across stakeholders. However, despite these challenges, these guidelines and principles remain crucial in fostering trust throughout the life-cycle of the AI system.

Given the increasing interest in AI's societal impact and its ethical alignment, TAI has emerged as a key concept in addressing critical concerns about AI governance. To bridge the existing research gap, this literature review synthesizes the current state of knowledge on TAI in information systems and related fields such as computer science. As the discourse on TAI continues to expand, it is imperative to consolidate existing research to provide a strong foundation for both scholars and practitioners. This synthesis aims to clarify key concepts, identify knowledge gaps and challenges, and highlight opportunities for future research. In this context, we formulate the following research question:

RQ: What is trustworthy artificial intelligence, and who are the stakeholders?

By addressing this question, the study contributes to a more comprehensive understanding of TAI, and highlights the importance of considering whose trust is impacted by AI systems, emphasizing the diverse stakeholders and contexts in which trust operates.

2. Background

2.1. Trust and Trustworthiness

Trust is relational and complex, involving at least two actors: one who trusts (the trustor) and one who is trusted (the trustee). In this relationship, the trustor relies on the trustee to carry out (or not carry out) a particular action [16]. Numerous definitions of trust exist across different disciplines, but Lukyanenko et al. [17] offer a general definition by generalising human mental

and physiological mechanisms to trust agents, such as AI agents. They define trust as "general trust is an information processing and behavioural process within a trusting agent that considers the properties of another system to control the extent and parameters of the interaction with this system." In the context of Information Systems, Mayer et al. [18] describe trust as "trust is the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." Duenser and Douglas [19] add that "trust relationships involving AI are socio-technical in nature, incorporating not only the AI itself but also people, laws, social norms, and institutions." Similarly, Siau and Wang [20] propose three perspectives on trust: (1) a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trusting beliefs); (2) the willingness of one party to depend on another in a risky situation (trusting intention); or (3) the combination of these elements." While closely related, trust and trustworthiness are distinct concepts. Trustworthiness, on the other hand, is a trait (characteristic), often confused with trust itself. Being trustworthy does not automatically create a trust relationship, nor does it guarantee that trust will be established [16]. The following section elaborates on the transition from trustworthy computing to TAI.

2.2. Trustworthy Computing to Trustworthy-AI

In the United States, 1999 report "Trust in Cyberspace" by the National Academies [21, 22] laid the foundations for what is now called "trustworthy computing" as a key area of research. Shortly after, the national science foundation (NSF) launched a series of initiatives to advance this field, starting with the "trusted computing program" in 2001, followed by the "cyber trust initiative" in 2004. By 2007, the trustworthy computing program was introduced, and in 2011, it evolved into the secure and trustworthy cyberspace initiative (SaTC). Spearheaded by the NSF's Computer and Information Science and Engineering Directorate, these efforts expanded trustworthy computing research beyond computer science into an interdisciplinary endeavor.

Industry has also played a central role in shaping the trustworthy computing landscape, with industry leaders like Microsoft at the forefront. In January 2002, Bill Gates issued his famous "Trustworthy Computing" memo [23], which marked a turning point for Microsoft and the broader tech industry. Gates' directive highlighted the urgent need for the development of trustworthy software and hardware products. The memo drew upon an internal Microsoft white paper that defined four key pillars of trustworthiness: security, privacy, reliability, and business integrity. This shift not only influenced Microsoft's approach but also set a precedent for the broader IT sector, solidifying the importance of trust as a core element of modern computing practices. According to Wing [22], trustworthy computing comprises a set of overlapping properties—reliability, safety, security, privacy, availability, and usability—applicable to hardware and software systems and their interactions with humans and the physical world. Wing [22] also emphasized that TAI systems require a more comprehensive set of properties than traditional computing systems. TAI encompasses not only reliability, security, privacy, and usability, but also additional properties such as probabilistic accuracy under uncertain conditions, fairness, robustness, accountability, and explainability.

2.3. Trustworthy AI

The term "artificial intelligence" was first conceived at a workshop at Dartmouth college in 1956 [24]. Since then, the field has undergone several waves of rapid progress [25]. Since the early 2010s, machine learning and deep learning have achieved groundbreaking advancements, with the pace of progress continuing to accelerate. These developments have fueled visions of a world enriched by intelligent agents enhancing individual's lives, organizations, and societies. However, AI is not a universal solution or a magic bullet. Like any other technology, it offers significant benefits while also introducing new ethical, legal, and social challenges that must be carefully addressed [26, 27]. The growing recognition of these challenges has led to the proliferation of AI frameworks, guidelines, and reports in recent years. Notably, two well-established public repositories of AI ethics guidelines include algorithmwatch [28] and linking AI principles (LAIP) [29]: Algorithmwatch, an organization that monitors the societal impact of digitalization, curated a collection of 167 AI ethics guidelines before ceasing new submissions in April 2024 [28]. Meanwhile, LAIP continues to compile ethical frameworks, with 115 proposals recorded as of April 2, 2025 [29]. Overall, these guidelines and recent research on trustworthiness [30] in the context of AI provide the key foundation for exploring the landscape of TAI in this review. These frameworks and guidelines aim to guide the design, development, and implementation of AI systems in ways that benefit individuals, businesses, and society while reinforcing human-centric values. However, different concepts and terminologies are often interpreted differently by various users and organizations. AI systems encompass both technical artefacts and human operators [31, 32] and not isolated; they are embedded within social contexts, forming complex socio-technical systems where society, technology, and organisation mutually shape and influence one another [33]. There are various definitions of TAI, however according to the European Commission's AI high-level expert group (HLEG), TAI must comply with legal, ethical, and technical standards while also being socially robust [8].

2.4. Stakeholders of Trustworthy AI (TAI)

Trust is a fundamental factor in the moral and ethical treatment of stakeholders. Hosmer describes trust as "the expectation by one person, group or firm of ethically justifiable behavior on the part of the other person, group or firm in a joint endeavor or economic exchange [34]." In the context of TAI systems, a diverse range of stakeholders, including designers, developers, AI/ML experts, data scientists, system engineers, project and product managers, regulators, funding organizations, auditors, and users of AI technology, among others, all play a role [35].

The responsibility for ensuring the development and use of TAI systems extends to individuals and organizations involved in their design, development, and maintenance, as well as legislative and regulatory bodies at both the national and international levels [35]. From a practical point of view, stakeholders directly influencing TAI system development include those responsible for designing, building, and maintaining the systems, as well as the organizations funding their creation. In contrast, those shaping ethical AI guidelines, such as public, private, and non-profit organizations, and those involved in developing legal and regulatory frameworks for AI, play an indirect role in shaping the behavior of AI systems. These actors must collaborate to establish rules and frameworks that address the impacts of TAI. However, as the ethical AI guidelines

gain further clarity, the influence of these stakeholders, across public, private, and non-profit sectors, will become more pronounced. Similarly, as national and international regulatory bodies develop more detailed legal obligations for TAI systems, their impact on the day-to-day practices of TAI development and deployment will continue to grow.

Despite the interest in developing TAI, there have been insufficient efforts to synthesize existing research on how it is thought about, as well as to evaluate how current frameworks address the needs and perspectives of various stakeholders (trustors) [36]. By broadening the scope to include diverse stakeholder perspectives, we can better address relational aspects of trust and ensure that TAI is designed to meet the needs of those it impacts.

3. Methodology

This literature review was conducted to explore current literature on trustworthy AI, its elements and relevance to various stakeholders. The chosen review methodology is inspired by the approach outlined by Webster & Watson [37], which offers valuable guidance for conducting a concept-centric analysis that synthesizing existing knowledge.

3.1. Paper Collection

The search process was guided by the following boundary conditions:

Inclusion Criteria:

- Research papers that discuss or theorize concepts related to TAI (interdisciplinary field).
- Papers on TAI and related ethical guidelines or principles.

Exclusion Criteria:

- Papers focused on purely computational or technological work.
- Papers addressing only specific elements of TAI rather than the whole concept or other terminologies except those that seem to be very close.

3.2. Search Process

We began the search process by selecting keywords, which we then used for literature searches. This was followed by evaluating the identified literature and selecting relevant articles. Finally, we conducted an in-depth analysis of the chosen literature.

3.2.1. Keyword Selection

We began by conducting a 'traditional' search with a broad query to identify fundamental keywords for discussing TAI's elements or dimensions, along with the relevant stakeholders involved. The reason for the broad query was to avoid limiting the number of hits and to identify the other concepts related to trustworthy AI. The search string, executed in the Scopus and Web of Science (WoS) databases ('trustworthy artificial intelligence' OR TAI OR ethics) AND (stakeholder*), helped identify articles addressing the ethical dimensions of TAI, including

synonyms of TAI, and the stakeholders involved to achieve TAI. The asterisk was used to ensure that different variations of word conjugations were included in the search (e.g., 'stakeholder,' and 'stakeholders'). Additionally, the 'literature review papers' discovered during the keyword search proved useful for finding related keywords.

3.2.2. Literature Search

We expanded the search query to include more than just trustworthiness-related terms, as certain terms can serve as synonyms for trustworthy AI. Additionally, we incorporated stakeholder terms to meet the requirements of this research work. Consequently, the query included:

Query: ("trustworthy artificial intelligence" OR TAI OR "responsible artificial intelligence" OR RAI OR "explainable artificial intelligence" OR XAI OR "ethical artificial intelligence" OR "trustworthy AI" OR "responsible AI" OR "explainable AI" OR "ethical AI" OR "lawful AI" OR "lawful act" OR "ethical guideline*" OR "ethical principle*" OR "ethical standard*") AND (stakeholder*).

Various combinations of keywords were employed to ensure relevancy in the search. The publications were collected from the digital repositories of the association for information systems AIS eLibrary (primarily to cover IS conference papers), Scopus database (using "https://litbaskets.io/" to identify relevant IS and related interdisciplinary journals), Web of Science (WoS), and the ACM Digital Library (Computing Literature).

The searches were performed on the title, abstract, and keywords of papers using the Boolean operators "OR" and "AND." The AIS eLibrary was searched due to its comprehensive coverage of the latest advancements in both practice and academia within IS research. Since IS is an interdisciplinary field that straddles other disciplines, it is often necessary to look not only within the IS discipline when reviewing and developing theory but outside the field (Webster & Watson, 2002). Therefore, the WoS, Scopus, and the ACM were also searched, which provide access to a diverse array of peer-reviewed literature, including academic journals, conference proceedings, and other scholarly documents across various academic fields.

An initial set of 896 records was compiled from four databases: 161 from the ACM Digital Library, 716 from Web of Science, 19 from Scopus (via Literature Baskets), and none from AIS eLibrary.

3.2.3. Literature Evaluation and Selection

The literature search was conducted with the aim to ensure comprehensive coverage of the existing literature. The articles are restricted to English language and only included peer-reviewed publications to ensure the research quality. The collection period for the AIS eLibrary, ACM, and WoS was from March 1, 2020, to March 31, 2025. For Scopus, due to date limitations, publications were collected from 2020 to 2025. The following items were excluded: articles that did not meet the publication date criteria, papers outside the search scope, books or book sections, articles shorter than six pages. Additionally, duplicates and triplicates were carefully avoided, and references were manually examined to uncover any additional articles or papers. The selection of papers was conducted in three steps. First, all fetched articles were screened to assess their eligibility based on the inclusion and exclusion criteria. In the

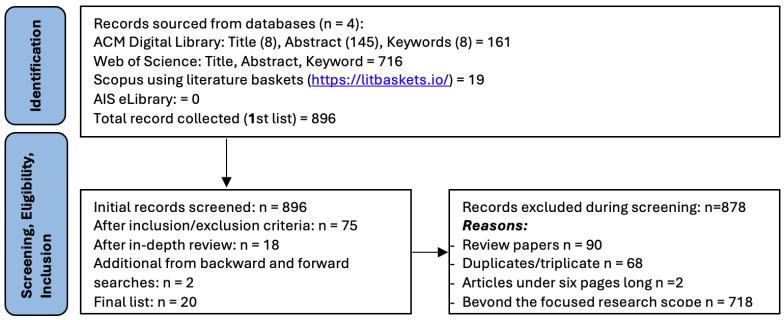

Figure 1: A custom visual PRISMA representation of the search process.

second step, the abstracts of the remaining publications were reviewed. In the third step, the full text of all remaining papers was thoroughly examined to identify those to be included. Finally, backward and forward searches were conducted using the papers identified as starting points, yielding two more relevant studies—one from each method. Throughout these steps, the inclusion and exclusion criteria were consistently applied. In total, twenty papers from the primary, backward, and forward sets were selected following in-depth review and filtering [35, 36, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55]. We used the PRISMA flow diagram [56] for clear and transparent reporting of the search process (see Figure 1).

In addition, sixteen frameworks, principles, or guidelines cited in the selected articles were analyzed, offering valuable insights into the components of TAI and related concepts. While some of these sources were published before 2020, it is important to emphasize that the primary goal of this research is not to provide an exhaustive review of all global AI policy frameworks, principles, or ethical guidelines. Rather, it aims to identify the key elements that define TAI and who are the stakeholders.

Table 1 categorizes the articles by terminology and publication year. Among the terms used, Explainable AI (XAI) is the most cited term (six articles), followed by Responsible AI (RAI) with four. Trustworthy AI (TAI), Ethical AI (EAI), and Human-Centric AI (HC-AI) appear in two articles each. Most publications appeared in 2024 (eight articles), followed by 2023 (five) and 2021 (four). One article was published in each of 2025, 2022, and 2020, totaling twenty articles.

Table 1
Selected articles classification by years and TAI related terminologies (on stakeholder perspective)

Year/TAI related terminology	TAI	RAI	EAI	XAI	HC-AI	Others	Total
2025	[36]	-	-	-	-	-	1
2024	-	[46]	[41]	[39, 45, 47]	-	[44, 49, 54]	8
2023	-	[48, 55]	-	[50]	[38, 43]	-	5
2022	-	[35]	-	-	-	-	1
2021	-	-	[52]	[42, 53]	-	[40]	4
2020	[51]	-	-	-	-	-	1
Total	2	4	2	6	2	4	20

4. Results

Table 2 presents a comprehensive overview of the fourteen selected AI frameworks, principles, or ethical guidelines, along with their associated themes. The leftmost column lists terminologies such as TAI, RAI, EAI, HC-AI, XAI, and beneficial AI (BAI) while the rightmost column indicates the number of principles associated with each framework.

Table 2
Overview of the fourteen randomly selected ethical frameworks

Terminologies	Framework	Theme	Principles Description
TAI	EC HLEG on AI (EU)	Lawful, ethical, & robust TAI	4 principles of TAI, encompassing 7 key requirements
TAI	OECD Principles on AI (Intl.)	Human rights, democratic values, & rule of law	5 principles for TAI
TAI	White House AI Principles (US)	Safe, ethical, & responsible development & deployment of AI	10 principles for TAI
TAI	G20 AI Principles	Human rights, diversity, sustainability, & global cooperation	5 principles for TAI
RAI	Chinese AI Principles	Development, governance, safety, reliability & controllability	8 principles for RAI
RAI	Facebook: RAI	Ethics considerations and societal impact	5 principles for RAI
RAI	Montreal Declaration	Fundamental interests of people and for the development of RAI	10 principles for RAI
EAI	Recommendation on EAI (Intl.). UNESCO	Human Rights	10 principles for EAI
EAI	IBM: Everyday EAI	Ethics embedded in AI design & development	5 principles for EAI
EAI	AI4People	assessment, development, incentivization, & support	5 principles for EAI, based on 20 action points
EAI	UK AI Code	Ethical AI code & development	5 principles for EAI
HCAI	Social Principles of HCAI	Social Principles of Humanity	7 principles of HCAI
XAI	NIST XAI principles (US)	Risk Management	4 Principles of XAI
BAI	Asilomar AI Principles	Research issues, ethics and values, and long-term issues	23 principles of beneficial AI

These TAI frameworks include AI HLEG (4 principles), OECD (5 principles), White House OSTP (10 principles), G20 AI principles (5 principles). RAI: Chinese AI Principles (8 principles), Facebook (5 principles), and Montreal Declaration (10 principles). EAI includes UNESCO's Recommendation (10 principles), IBM's Everyday EAI (5 principles), AI4People (5 principles), and UK AI Code (5 principles). HC-AI is represented by Japan's social principles (7 principles),

XAI includes NIST's XAI framework (4 principles), while beneficial AI (BAI) includes Asilomar AI Principles (23 principles). These frameworks were further validated using two well-established public repositories of AI ethics guidelines: AlgorithmWatch [28] and the LAIP repository [29].

Table 3
Overview of the core TAI and related terminologies and associated elements

Guidelines/Principles → **Socio-technical Elements** ↓ **Abbreviations:** Trustworthy AI - TAI Responsible AI - RAI Ethical AI - EAI Intelligent Analysis & IS - IAIS Human-Centric AI - HCAI Governance Principles for the New-Generation AI - GPNGAI European Comission - EU Organisation for Economic-Cooperation and Development - OECD	EC Guidelines for TAI - 2019	OECD. AI Principles - 2019	Fraunhofer (Germany)–IAIS - TAI Principles - 2020	G20 TAI Principles - 2019	UNESCO: Recommendation on EAI - 2021	Australia: EAI Framework - 2019	Japan: Social Principles of Human Centric HCAI - 2019	China: GPNGAI-Governance Principles for the New Generation AI: RAI - 2019	IBM: Everyday EAI - 2019	Facebook: RAI - 2021	Frequency of Mentions
Human agency and oversight	x		x		x			x		x	5
Human-centered values		x	x	x	x	x	x		x	x	8
Societal & environmental wellbeing	x	x	x	x	x	x		x			7
Contestability						x					1
Innovation and Solutions		x	x				x	x			4
Literacy / Education					x						1
Collabration/ Muti-stake- holder/Adaptive Goveranance					x					x	2
Inclusive growth/Sustainability		x	x	x	x		x	x		x	7
Cultural/Socital value sentitive			x						x		2
Technical robustness & safety	x	x	x	x	x	x		x		x	8
Privacy and data governance	x	x	x		x	x	x	x	x	x	9
Diversity, non-discrimination, & fairness /& democracy	x	x			x	x	x	x	x	x	8
Transparency	x	x	x	x	x	x	x	x		x	9
Accountability	x	x	x	x	x	x	x	x	x	x	10
Robustness		x	x	x						x	4
Fairness /& Justice		x	x	x		x	x	x			6
Explainability		x	x	x	x			x	x		6
Reliablity			x			x		x			3
Security		x	x	x	x		x	x			6
Proportionality					x						1
Harmony /& Friendliness			x					x			2
Ratio of partially added ethical elements to the original count in the guideline.	7/7	13/5	16/7	10/5	14/10	10/8	9/7	14/8	6/5	10/5	

Table 3 presents a comprehensive overview of the ten selected AI frameworks, principles,

or ethical guidelines, along with their associated elements. In the final column of Table 3, the frequency of each element mentioned across different frameworks having various themes. The last row of Table 3 displays the ratio of `partially included ethical elements`[1] relative to the total number of principles/requirements outlined in the original policy frameworks.

5. Discussion

In this section, we discuss the results, including key elements of TAI, the stakeholders involved in achieving TAI, and identify areas that have not been sufficiently explored. It also proposes the future directions for researchers and practitioners to advance the development of TAI. The discussion covers AI-community fragmentation, TAI and its key elements, stakeholders involved, and the limitations of the current approach.

5.1. AI - Community Fragmentation

AI, like any technology, offers significant benefits but also introduces new ethical, legal, and social challenges [26]. In response, numerous guidelines, frameworks, or guiding principles related to TAI and its related concepts have emerged in recent years. These frameworks often propose varying numbers of principles, requirements, or recommendations. Notably, differences arise not only between distinct AI-related terminologies but also within the same terminology as defined by different sources (see Table 2). As a result, the landscape of AI governance has become increasingly complex, with various overlapping and sometimes inconsistent attributes expected from AI systems. Terms like 'AI Safety,' 'Fairness in AI,' 'Secure AI,' 'Explainable AI,' 'Transparent AI,' 'Responsible AI,' 'Trustworthy AI,' 'Interpretable AI,' 'Robust AI,' 'Ethical AI,' 'Accountable AI,' 'Resilient AI,' 'Reliable AI,' 'Black-box AI,' 'Privacy-enhanced AI,' and 'Federated AI' reflect this diversity of perspectives in this space [57]. This diversity can complicate understanding and create the impression that AI governance communities are fragmented, and that the concept of TAI lacks coherence and a unified definition. Nevertheless, despite these terminological differences, these initiatives share a common objective: to enhance the benefits of AI while mitigating its associated risks and harms. Ultimately, these overlapping concepts reflect both multidisciplinary and interdisciplinary efforts to guide the development and deployment of TAI systems, guided by experts from various disciplines. Although many definitions of TAI have been proposed across various scenarios and themes, there is still no universally accepted definition of TAI. The following are some of the most commonly used TAI-related terms and their definitions, developed by experts from different disciplines and in different contexts.

Trustworthy AI: As defined by the european commission's high-level expert group (HLEG) on AI, trustworthy AI must adhere to legal, ethical, and technical standards while also being socially robust [8]. It concerns not only the trustworthiness of the AI system itself but also comprises the trustworthiness of all processes and actors involved in the system's life cycle.

Responsible AI: According to the world economic forum (WEF), responsible AI refers to the practice of designing, building, and deploying AI in a manner that empowers people

[1]Here partially included elements refer to cases where certain elements within specific frameworks serve as sub-elements under different themes, often interpreted with slight variations.

and businesses while ensuring fair impacts on customers and society [58]. This approach enables companies to engender trust and scale AI with confidence [58]. Arrieta et al. further define responsible AI as a methodology for the implementation of AI methods in real world organization with fairness, model explainability and accountability at its core [59].

Explainable AI: As noted by defense advanced research projects agency (DARPA), explainable AI aims to produce more explainable models, that maintain high learning performance (prediction accuracy) while enabling human users to understand, appropriately trust, and effectively manage AI systems [15].

Ethical AI, as defined by the european commission's high-level expert group (HLEG) on AI, ethical AI refers to the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, moral principles and related core values [8]. It is the second of the three core elements (lawful, ethical, and robust) necessary for achieving TAI according to the EC's HLEG on AI [8].

Despite differing terminologies, these initiatives share the common goal of maximizing AI's benefits while mitigating risks and potential harm.

5.2. Key Elements of TAI

The goal of this review is to enhance understanding of TAI, identify its key elements tailored to various contexts or themes, and identify the associated stakeholders. To explore the key elements of TAI, we began by reviewing ten AI policy frameworks, guiding principles, or guidelines (see Table 3), tailored to different thematic areas (some listed in Table 2) and introduced by various global organizations. These documents define the essential elements that AI systems must possess to be considered trustworthy. Categorizing these socio-technical elements proved challenging, as elements identified as primary in one framework often appeared as sub-elements in another. Moreover, these elements were often organized under different context or themes, sometimes with slight variations in interpretation. Considerable effort was invested in organizing these elements and assessing their frequency of occurrence across the frameworks. Among the most frequently cited elements—appearing in at least eight of the ten frameworks—were:

Accountability: A relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences [60]. **Transparency:** It refers to the need to explain, interpret, and reproduce its decisions [61]. **Privacy and data governance:** It emphasizes protecting data privacy, integrity, and quality, while ensuring individual's rights to access their data. [62]. **Technical robustness and safety:** This means, AI systems should be technically robust and perform as expected by the users [8]. The system should recover safely from failures, handle errors throughout the AI lifecycle, resist external attacks, and produce reproducible results. **Diversity, non-discrimination, and fairness:** This means, AI systems should ensure fairness and avoid direct or indirect discrimination against any societal group, regardless of socio-economic factors [63]. **Human-centered values:** This means that the use of AI must not infringe upon the fundamental human rights guaranteed by the Constitution and international standards [64]. AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights [8]. Other important elements—including societal and environmental well-being, inclusive growth

and sustainability, fairness and justice, explainability, and security—were cited in at least six frameworks. Additionally, human agency and oversight appeared in five frameworks. In total, we identified twenty-one distinct categorized elements across the selected frameworks. It is important to note that some categories include multiple elements. This is due to the fact that different frameworks use varying terminologies and structure their categories on diverse themes. Consequently, the intended stakeholders may differ across frameworks, potentially leading to confusion when interpreting elements associated to different contexts or thematic priorities. Even elements cited less frequently still contribute meaningfully to the essential requirements of TAI.

However, Jobin et al. [7] reviewed 84 ethical AI documents and found that while no single principle/element appeared in all, elements like transparency, fairness, responsibility, non-maleficence, and privacy were present in over half. The consensus is fragile due to inconsistent terminology and lack of legal grounding—apart from the European AI Act (passed, but is still under continued development)–leaving room for countries or companies to adopt alternative principles for convenience or competitive advantage. Still, this emerging common ground provides a useful foundation for aligning stakeholder expectations and guiding co-design in the interdisciplinary field of AI, although challenges remain.

TAI elements are often interrelated; for instance, transparency supports accountability, and fairness contributes to societal well-being. However, balancing these values in practice frequently requires design trade-offs. A common tension arises between explainability and performance, where more interpretable models—such as decision trees—may underperform compared to more complex, deep neural network models. Similarly, efforts to enhance fairness, such as mitigating bias, can sometimes lead to reduced accuracy for the majority class. The tension between privacy and utility is also well-documented; for example, implementing differential privacy may protect individual data but reduce the utility of datasets. Overall, these findings highlight the core elements commonly emphasized in TAI discourse, suggesting that existing frameworks may structure these elements differently depending on the thematic focus they are developed around and the stakeholders involved. This observation can be seen as a contribution to the field, offering a batter understanding of the evolving landscape of TAI.

5.3. The Intended Stakeholders for TAI

Duenser and Douglas [19] stated that: “Trust relationships involving AI are socio-technical in nature, incorporating not only the AI itself but also people, laws, social norms, and institutions.” Viewing AI as part of a socio-technical system is essential, as the roles of people (the stakeholders) involved, those who design, develop, deploy, governance, guide and use the technology, are as critical as the technology itself in establishing its trustworthiness.

Viewing TAI through a stakeholder lens enhances our socioeconomic understanding of the primary challenges in achieving TAI. It emphasizes the interdependency among various stakeholders, the themes for which TAI systems are developed, the technology itself, and the social contexts in which TAI systems are developed and deployed. This perspective shifts the focus from seeing AI merely as a technical artifact to recognizing TAI as a system deeply embedded within human, organizational, and societal structures. Consequently, AI technologies must be understood within their socio-technical contexts, with a particular emphasis on stakeholders,

rather than shareholders — both during their development and application.

From a practical perspective, the primary stakeholders directly influencing the development of TAI systems include those involved in designing, developing, and maintaining these systems, as well as the organizations that fund their creation. Those shaping the ethical AI guidelines, such as public, private, and non-profit organizations, and those involved in developing legal and regulatory frameworks for AI, play an indirect role in shaping the behavior of AI systems. Additionally, the ecosystem of TAI stakeholders extends to include not only the users of these systems but also those affected by their deployment. TAI, therefore, functions as a socio-technical system, encompasses different stakeholders including developers, its users, the technology, and the institutions that govern the interactions among different stakeholders. The term "institutions that shape these interactions" refers to the formal and informal organizations, policies, regulations, cultural norms, and frameworks that shape how TAI technologies are developed, deployed, and used. These can include governmental bodies, regulatory agencies, industry standards organizations, educational institutions, and even societal norms or ethical guidelines. These institutions establish rules (such as EU AI act), provide oversight, and set expectations that influence how stakeholders engage with AI systems. Effective collaboration among these actors is crucial for building sustainable TAI systems and for continuous updating and implementing rules and frameworks that address the broader impacts of TAI.

It is evident that various AI policy frameworks, principles, or guidelines have been developed with varying thematic considerations in mind, and none can be considered perfect or complete for every scenario. While, these documents present high-level objectives for TAI systems and the underlying science and technology, they do not delve into specific technical implementations. Moreover, the frameworks show significant overlap, with widespread agreement that AI should promote the common good, avoid causing harm or violating rights, and uphold fundamental human values.

Different stakeholders often have conflicting priorities. Such as, developers often prioritize optimizing model performance, whereas regulators and users emphasize concerns about systemic biases. Organizations may resist transparency due to proprietary algorithms, while civil society groups and users increasingly demand explainability and accountability. In domains such as health or mobility, researcher's desire for open data access may conflict with individual's rights to privacy and consent. Moreover, the application of international AI principles may clash with local legal systems, cultural values, or societal norms, creating friction across geopolitical and social contexts. To address these conflicts, we believe all stakeholders must come together to engage in dialogue, negotiate trade-offs, and collaboratively develop solutions.

As ethical AI guidelines continue to evolve and become more defined, the influence of stakeholders across the public, private, and non-profit sectors is expected to grow significantly. Similarly, national and international regulatory bodies will play an increasingly important role as they introduce more detailed legal obligations with stakeholder responsibilities, further shaping the daily practices involved in the development and deployment of TAI systems.

5.4. Limitations

This study limited to four databases—AIS eLibrary, scopus (via litbaskets), web of science, and ACM digital Library— may not have been exhaustive, but it is expected that these studies will

contribute sufficient knowledge, aiding both novice and experienced researchers within the studied subject. Furthermore, this study does not cover prior research on TAI-related topics published before March 1, 2020, except for those in scopus databases, due to specific date constraints (covering 2020–2025).

5.5. Conclusion

This study was motivated by the growing importance of TAI and the lack of a comprehensive understanding of its key elements and the stakeholders involved. In response, we conducted a scoping review to examine how TAI is conceptualized, what core TAI elements are emphasized, and which stakeholders are involved.

Through thematic analysis, we identified twenty-one distinct elements. Among the most frequently cited elements—appearing in at least eight of the ten frameworks—were accountability, privacy and data governance, transparency, human-centered values, technical robustness and safety, and diversity, non-discrimination, fairness, and democracy. Other important elements—including societal and environmental well-being, inclusive growth and sustainability, fairness and justice, explainability, and security—were cited in at least six frameworks. Additionally, human agency and oversight appeared in five frameworks. These elements are drawn from ethical frameworks developed around diverse themes, reflecting different contextual priorities and stakeholder perspectives. The diversity in structure and slight variations in interpretation reveal the challenges of consistently categorizing elements and highlight the potential ambiguities in correctly identifying elements and their corresponding stakeholder relevance.

Our findings underscore that TAI is not merely a technical challenge but a socio-technical endeavor involving a diverse ecosystem of stakeholders. These include developers, designers, funding organizations, policymakers, regulatory agencies, users, and affected communities. The study shows that ethical principles are often framed differently depending on institutional priorities, cultural norms, and sectoral interests—resulting in distinct thematic orientations across frameworks. Consequently, understanding TAI elements and stakeholders within their specific contexts is essential for effectively interpreting and applying these frameworks. Viewing AI through a stakeholder lens enhances our socioeconomic understanding of the key challenges in achieving TAI. It highlights the inter-dependencies among stakeholders, the purposes for which AI systems are developed, the technologies themselves, and the social contexts in which they are implemented.

This observation can be seen as a contribution to the field, offering a batter understanding of the evolving landscape of TAI, and the stockholders involved achieving TAI. We encourage further exploration, validation of these findings, and the incorporation of new insights to advance the ongoing discourse on TAI.

Acknowledgments

This research was supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS), funded by the Wallenberg Foundations, Sweden.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-4 (OpenAI) to check grammar and spelling. The author(s) subsequently reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Anshari, M. Hamdan, N. Ahmad, E. Ali, Public service delivery, artificial intelligence and the sustainable development goals: trends, evidence and complexities, Journal of Science and Technology Policy Management 16 (2025) 163–181.

[2] A. B. Rashid, A. K. Kausik, Ai revolutionizing industries worldwide: A comprehensive overview of its diverse applications, Hybrid Advances (2024) 100277.

[3] S. J. Mikhaylov, M. Esteve, A. Campion, Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration, Philosophical transactions of the royal society a: mathematical, physical and engineering sciences 376 (2018) 20170357.

[4] V. Dignum, Responsible artificial intelligence: how to develop and use AI in a responsible way, volume 2156, Springer, 2019.

[5] G. Mentzas, M. Fikardos, K. Lepenioti, D. Apostolou, Exploring the landscape of trustworthy artificial intelligence: Status and challenges, Intelligent Decision Technologies 18 (2024) 837–854.

[6] M. Khaleel, A. Jebrel, D. M. Shwehdy, Artificial intelligence in computer science: https://doi.org/10.5281/zenodo. 10937515, Int. J. Electr. Eng. and Sustain. (2024) 01–21.

[7] A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature machine intelligence 1 (2019) 389–399.

[8] H. AI, High-level expert group on artificial intelligence, Ethics guidelines for trustworthy AI 6 (2019).

[9] OECD, Oecd principles on ai: Recommendation of the council on artificial intelligence, https://www.oecd.org/en/topics/ai-principles.html, 2019. Accessed: 18 March 2025.

[10] R. T. Vought, Guidance for Regulation of Artificial Intelligence Applications, Technical Report, US White House, Washington, DC, 2020. URL: https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf, accessed: 18 March 2025.

[11] GPNGAI - Chinese National Governance Committee for the New Generation Artificial Intelligence, Governance principles for the new generation artificial intelligence–developing responsible artificial intelligence, https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html, 2019. Accessed: 18-Mar-2025.

[12] A. P. Team, Artificial intelligence measurement and evaluation at the national institute of standards and technology, National Institute of Standards and Technology (2021).

[13] UNESCO - United Nations Educational, Scientific and Cultural Organization, Recommendations on the ethics of artificial intelligence, 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137, accessed: 18 March 2025.

[14] International Organization for Standardization, ISO 24028:2020 information technology–

artificial intelligence–overview of trustworthiness in ai, https://www.iso.org/standard/77608.html, 2020. Standard.
[15] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI magazine 40 (2019) 44–58.
[16] F. Gille, A. Jobin, M. Ienca, What we talk about when we talk about trust: theory of trust for ai in healthcare, Intelligence-Based Medicine 1 (2020) 100001.
[17] R. Lukyanenko, W. Maass, V. C. Storey, Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities, Electronic Markets 32 (2022) 1993–2020.
[18] R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of management review 20 (1995) 709–734.
[19] A. Duenser, D. M. Douglas, Whom to trust, how and why: untangling artificial intelligence ethics principles, trustworthiness, and trust, IEEE Intelligent Systems 38 (2023) 19–26.
[20] K. Siau, W. Wang, Building trust in artificial intelligence, machine learning, and robotics, Cutter business technology journal 31 (2018) 47.
[21] N. R. Council, et al., Trust in cyberspace, National Academies Press, 1999.
[22] J. M. Wing, Trustworthy ai, Communications of the ACM 64 (2021) 64–71.
[23] B. Gates, Trustworthy computing, Wired News, Jan 17 (2002).
[24] J. McCarthy, M. L. Minsky, N. Rochester, C. E. Shannon, A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955, AI magazine 27 (2006) 12–12.
[25] M. Haenlein, A. Kaplan, A brief history of artificial intelligence: On the past, present, and future of artificial intelligence, California management review 61 (2019) 5–14.
[26] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations, Minds and machines 28 (2018) 689–707.
[27] L. Floridi, Establishing the rules for building trustworthy ai, Ethics, Governance, and Policies in Artificial Intelligence (2021) 41–45.
[28] A. Watch, The ai ethics guidelines global inventory, 2019. URL: https://inventory.algorithmwatch.org/, accessed: 2025-03-18.
[29] Laip - linking artificial intelligence principles, https://www.linking-ai-principles.org/, 2018. Accessed: 2025-03-18.
[30] M. M. Ferdaus, M. Abdelguerfi, E. Ioup, K. N. Niles, K. Pathak, S. Sloan, Towards trustworthy ai: A review of ethical and robust large language models, arXiv preprint arXiv:2407.13934 (2024).
[31] I. Van de Poel, Embedding values in artificial intelligence (ai) systems, Minds and machines 30 (2020) 385–409.
[32] P. Di Maio, Towards a metamodel to support the joint optimization of socio technical systems, Systems 2 (2014) 273–296.
[33] E. L. Trist, The evolution of socio-technical systems, volume 2, Ontario Quality of Working Life Centre Toronto, 1981.
[34] L. T. Hosmer, Trust: The connecting link between organizational theory and philosophical

ethics, Academy of management Review 20 (1995) 379–403.
[35] A. Deshpande, H. Sharp, Responsible ai systems: who are the stakeholders?, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 227–236.
[36] C. D. Wirz, J. L. Demuth, A. Bostrom, M. G. Cains, I. Ebert-Uphoff, D. J. Gagne II, A. Schumacher, A. McGovern, D. Madlambayan, (re) conceptualizing trustworthy ai: A foundation for change, Artificial Intelligence (2025) 104309.
[37] J. Webster, R. T. Watson, Analyzing the past to prepare for the future: Writing a literature review, MIS quarterly (2002) xiii–xxiii.
[38] U. Wilkens, D. Lupp, V. Langholf, Configurations of human-centered ai at work: seven actor-structure engagements in organizations, Frontiers in Artificial Intelligence 6 (2023) 1272159.
[39] Z. Wang, C. Huang, X. Yao, A roadmap of explainable artificial intelligence: Explain to whom, when, what and how?, ACM Transactions on Autonomous and Adaptive Systems 19 (2024) 1–40.
[40] N. M. Pless, A. Sengupta, M. A. Wheeler, T. Maak, Responsible leadership and the reflective ceo: Resolving stakeholder conflict by imagining what could be done, Journal of Business Ethics (2021) 1–25.
[41] T. K. Mitchell, M. Popa, R. E. Ashcroft, S. Prasad, A. Sharp, C. Carnforth, M. Turner, A. Khalil, N. Fenwick, S. Leven, et al., Balancing key stakeholder priorities and ethical principles to design a trial comparing intervention or expectant management for early-onset selective fetal growth restriction in monochorionic twin pregnancy: Fern qualitative study, BMJ open 14 (2024) e080488.
[42] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research, Artificial intelligence 296 (2021) 103473.
[43] M. Langer, C. J. König, Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management, Human Resource Management Review 33 (2023) 100881.
[44] M. Kinney, M. Anastasiadou, M. Naranjo-Zolotov, V. Santos, Expectation management in ai: A framework for understanding stakeholder trust and acceptance of artificial intelligence systems, Heliyon 10 (2024).
[45] M. Kim, S. Kim, J. Kim, T.-J. Song, Y. Kim, Do stakeholder needs differ?-designing stakeholder-tailored explainable artificial intelligence (xai) interfaces, International Journal of Human-Computer Studies 181 (2024) 103160.
[46] A. Kawakami, D. Wilkinson, A. Chouldechova, Do responsible ai artifacts advance stakeholder goals? four key barriers perceived by legal and civil stakeholders, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, volume 7, 2024, pp. 670–682.
[47] S. Haas, K. Hegestweiler, M. Rapp, M. Muschalik, E. Hüllermeier, Stakeholder-centric explanations for black-box decisions: an xai process model and its application to automotive goodwill assessments, Frontiers in Artificial Intelligence 7 (2024) 1471208.
[48] M. Marzouk, C. Zitoun, O. Belghith, S. Skhiri, The building blocks of a responsible artificial intelligence practice: An outlook on the current landscape, IEEE Intelligent Systems 38

(2023) 9–18.
[49] S. Gupta, R. Jaiswal, How can we improve ai competencies for tomorrow's leaders: Insights from multi-stakeholders' interaction, The International Journal of Management Education 22 (2024) 101070.
[50] R. R. Hoffman, S. T. Mueller, G. Klein, M. Jalaeian, C. Tate, Explainable ai: roles and stakeholders, desirements and challenges, Frontiers in Computer Science 5 (2023) 1117848.
[51] R. Madhavan, J. A. Kerr, A. R. Corcos, B. P. Isaacoff, Toward trustworthy and responsible artificial intelligence policy development, IEEE Intelligent Systems 35 (2020) 103–108.
[52] S. Fukuda-Parr, E. Gibbons, Emerging consensus on 'ethical ai': Human rights critique of stakeholder guidelines, Global Policy 12 (2021) 32–44.
[53] A. Kasirzadeh, Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence, arXiv preprint arXiv:2103.00752 (2021).
[54] C. Figueras, C. Rossitto, T. Cerratto Pargman, Doing responsibilities with automated grading systems: An empirical multi-stakeholder exploration, in: Proceedings of the 13th Nordic Conference on Human-Computer Interaction, 2024, pp. 1–13.
[55] D. Domínguez Figaredo, J. Stoyanovich, Responsible ai literacy: A stakeholder-first approach, Big Data & Society 10 (2023) 20539517231219958.
[56] N. R. Haddaway, M. J. Page, C. C. Pritchard, L. A. McGuinness, Prisma2020: An r package and shiny app for producing prisma 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis, Campbell systematic reviews 18 (2022) e1230.
[57] M. M. Nasr-Azadani, J.-L. Chatelain, The journey to trustworthy ai-part 1: Pursuit of pragmatic frameworks, arXiv preprint arXiv:2403.15457 (2024).
[58] World Economic Forum, 4 steps to developing responsible ai, 2019. URL: https://www.weforum.org/agenda/2019/06/4-steps-to-developing-responsible-ai/, accessed: 2025-04-06.
[59] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.
[60] M. Bovens, Analysing and assessing accountability: A conceptual framework 1, European law journal 13 (2007) 447–468.
[61] V. Dignum, Responsible artificial intelligence: designing ai for human values (2017).
[62] R. V. Zicari, J. Brusseau, S. N. Blomberg, H. C. Christensen, M. Coffee, M. B. Ganapini, S. Gerke, T. K. Gilbert, E. Hickman, E. Hildt, et al., On assessing trustworthy ai in healthcare. machine learning as a supportive tool to recognize cardiac arrest in emergency calls, Frontiers in Human Dynamics 3 (2021) 673104.
[63] A. W. Flores, K. Bechtel, C. T. Lowenkamp, False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks, Fed. Probation 80 (2016) 38.
[64] Government of Japan, Social Principles of Human-Centric AI, Technical Report, 2019. URL: https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf.