Human-Intelligent Systems Integration (2025) 7:373–390
https://doi.org/10.1007/s42454-025-00068-w

ORIGINAL PAPER

E-LENS: user requirement–oriented AI ethics assurance

Jianlong Zhou[1] · **Fang Chen**[1]

Received: 7 March 2025 / Accepted: 7 August 2025 / Published online: 28 August 2025
© The Author(s) 2025

Abstract
Despite the much proliferation of AI ethical principles in recent years, there is a challenge of assuring AI ethics with current AI ethics frameworks in real-world applications. While system safety has emerged as a distinct discipline for a long time, originated from safety concerns in early aircraft manufacturing, the safety assurance is now an indispensable component in safety critical domains. Motivated by the assurance approaches for safety–critical systems such as aviation, this paper introduces the concept of AI ethics assurance cases into the AI ethics assurance. Three pillars of user requirements, evidence, and validation are proposed as key components and integrated into AI ethics assurance cases for a new approach of user requirement–oriented AI ethics assurance. The user requirement–oriented AI ethics assurance case is set up based on three pillars and hazard analysis methods used in the safety assurance of safety–critical systems. This paper also proposes a platform named Ethical-Lens (E-LENS) to implement the user requirement–oriented AI ethics assurance approach. The proposed user requirement–based E-LENS platform is then applied to assure AI ethics of an AI-driven human resource shortlisting system as a case study to show the effectiveness of the proposed approach.

Keywords AI ethics · User requirements · Ethics assurance

1 Introduction

Artificial Intelligence (AI) is considered a computer system with at least one of capabilities of perception (e.g. tactile, textual, visual, and audio), decision-making (e.g. financial investment), prediction (e.g. crop yield and health risk), pattern recognition (e.g. anomaly identification), interactive communication (e.g. chat bots or social robots), and many other intelligent functions (Vinuesa et al. 2020; Zhou and Chen 2023). These impressive capabilities allow AI perform tasks intelligently comparable with the human intelligence and even much better than humans, which motivates wide applications of AI in various domains such as agriculture, public services, healthcare, education, and others to enhance management performance and promote productivity. AI has been becoming an indispensable component in people's everyday life. However, AI systems are fast, complex, different from humans, and may change via learning, resulting in the societal impacts at a large scale. Furthermore, encased in a black box, contemporary AI solutions often obscure their inner workings in a way. And, ultimately, it is difficult to have trust and confidence in something we do not truly, genuinely, understand. Their behaviour can also be difficult to monitor, validate, predict, and explain. Moreover, AI needs incredible volumes of complex interconnected data (including personal data) for model building, while biases, errors, and bad habits in data are being absorbed into most AI systems. The use of personal data also puts people at risk. As a result, AI faces significant ethical challenges ranging from data privacy, explainability, ownership, and privacy, trust, fairness to accountability and others. In short, two core questions need to be answered on these ethical concerns:

- What is the right thing to do by AI?
- How do we know for whom and whether AI is "good"?

In order to mitigate ethical concerns of AI, various AI ethical frameworks and guidelines have been proposed by national and international organisations as well as non-profit organisations and professional associations (Zhou et al. 2020). For example, the AI ethical principles identified by

✉ Jianlong Zhou
jianlong.zhou@uts.edu.au

Fang Chen
fang.chen@uts.edu.au

[1] Human-Centered AI Lab, Data Science Institute, University of Technology Sydney, Sydney, NSW 2007, Australia

Data61 in Australia include human-centred values; human, social, and environmental wellbeing; fairness; privacy protection and security; transparency and explainability; reliability and safety; contestability; and accountability (Dawson et al. 2019). IEEE's AI ethical principles include wellbeing, human rights, transparency, accountability, data agency, effectiveness, awareness of misuse, and competence (Chatila and Havens 2019). It was found that no single ethical principle is explicitly agreed by all published ethical guidelines. However, it shows a convergence around the principles of beneficence, responsibility, transparency, justice and fairness, non-maleficence, privacy, trust, sustainability, dignity, freedom and autonomy, and solidarity in the global policy landscape (Jobin et al. 2019). Furthermore, organisations identified mandatory AI ethical principles based on those convergent ethical principles: community benefit, transparency, fairness, privacy and security, and accountability (Digital.NSW 2024).

Despite the much proliferation of AI ethical principles in recent years, there is a challenge of assuring AI ethics with current AI ethics frameworks in real-world applications (Zhou and Chen 2023). The current frameworks of AI ethics are usually abstract and provide limited approaches for developers and researchers of algorithms to apply them in practical applications. For example, it was found that 75% of ethical frameworks only contain general principles with very little details, and more than 80% of them offer no or very little information on practical applications (Bélisle-Pipon, et al., 2023); Zicari et al. 2022). Checklist-style questionnaire is one of widely accepted tools for the operationalisation of AI ethical principles. The *Assessment list for trustworthy artificial intelligence (ALTAI)* published by the HighLevel Expert Group on Artificial Intelligence (HLEG-AI) of European Commission in 2020 is one of checklist-style questionnaire examples (Ala-Pietil¨a et al. 2020). Han and Choi (2022) devise a checklist to review trustworthiness of AI based on ethical principles of explainability, fairness, robustness, safety, and transparency. However, the checklist questions are mostly simple yes/no validations and there are no further detailed insights/suggestions, which are not enough for computational models. The documentation is another intervention that has been adapted and developed to communicate key facts about machine learning (ML) systems. Datasheets (Gebru et al. 2021), FactSheets (Arnold et al. 2019), and Method Cards (Adkins et al. 2022) are documentation tools proposed to provide key details about datasets, models, and other details about AI development. Furthermore, frameworks are developed by providing an integrated process to assure AI ethics. For example, *Z*-inspection (Zicari et al. 2021) outlines a process with three phases of setup, assess, and resolve phases to assess trustworthy AI. However, these endeavours also often only focus on general questions and do not consider requirements of specific domains. They pay little attention to user requirements on AI solutions, which is the key for both AI solutions and AI ethics.

Moreover, system safety has emerged as a distinct discipline for a long time, originated from safety concerns in early aircraft manufacturing (Rinehart et al 2015). The safety assurance is now an indispensable component in safety critical domains such as nuclear, aviation, energy, rail, and finance (Rinehart et al. 2015), which is characterised by high levels of safety criticality, and technical and operational complexity. For example, Sharifi, Amyot, Mylopoulos, McLaughlin, and Feodoroff (2022) use requirements engineering techniques and standards, which are common in safety–critical sectors, to set up a set of requirement-based guidelines to build a rigorous certification for FinTech systems. Goal modelling, systemic safety assessment, and assurance case development are used in these guidelines. Safety cases try to provide an established approach for validating the safety of systems through an explicit argument supported by compelling evidence (Picardi, Paterson, Hawkins, Calinescu, and Habli 2020).

Motivated by the assurance approaches for safety–critical systems such as aviation (Rinehart et al. 2015), this paper introduces the concept of assurance cases into AI ethics assurance. Three pillars of user requirements, evidence, and validation (REV) are proposed and integrated into AI ethics assurance cases for a new approach of user requirement–oriented AI ethics assurance. User requirements on AI ethics are user's specific interest in ethical concerns and they are the basis for the assurance. The clear identification of user requirements on ethics allows efficient assurance of targeted ethical issues instead of non-focused general ethical problems. Furthermore, the AI ethics in real-world applications must be assured with enough evidence that AI models are used ethically. Such evidence must be sufficient to explain why the AI system can be trusted ethically for its applications. The evidence is then validated and verified by experts in AI ethics and/or regulators so that AI systems can be endorsed and used ethically. Based on these three pillars, the AI ethics assurance case is set up using system theoretic process analysis (STPA) (Leveson 2016) which is a hazard analysis method widely used in the safety assurance of safety–critical systems. This paper also proposes a platform named Ethical-Lens (E-LENS) to implement the user requirement–oriented AI ethics assurance approach.

The primary contributions of this paper include:

- The concept of assurance case widely used in safety–critical systems is introduced into AI ethics assurance to validate AI ethical principles.
- A new approach of user requirement–oriented AI ethics assurance is proposed by integrating three pillars of

user requirements, evidence, and validation into AI ethics assurance cases.
- A new platform of E-LENS is proposed to implement the user requirement–oriented AI ethics assurance approach.

2 Related work

2.1 AI ethics operationalisation

Han and Choi (2022) formulate the considerations for validating a trustworthy AI and devise a checklist to review trustworthiness of AI. The checklist items are based on ethical principles of explainability, fairness, robustness, safety, and transparency throughout the AI life cycle. The High-Level Expert Group on Artificial Intelligence (HLEG-AI) of European Commission published the *Assessment List for Trustworthy Artificial Intelligence (ALTAI)* in 2020, which represents a significant step in terms of AI ethics assurance. The ALTAI checklist tool enables users to conduct assessments of AI systems about the governance to the ethical principles (AlaPietilä et al. 2020). Some studies have used ALTAI in examining specific contexts, such as how users should achieve ethical AI, or how the ALTAI list can be applied in driver-assistance systems (Borg et al. 2021; Gardner et al. 2022). Radclyffe et al. (2023) review the pros and cons of the ALTAI checklist tool. The major strength lies in its capability to implement the HLEG-AI guideline into an objective tool that can measure the assurance results. The weakness shows that the ALTAI checklist tool does not provide guidance for an organisation's level of development on AI ethics. ALTAI also lacks actionable recommendations, particularly aimed at industry developments and practices on AI technologies (Radclyffe et al. 2023). It also lacks quantitative evaluations of ethics on some ethical principles such as explainability and fairness. Furthermore, the checklist questions are mostly only simple yes/no validations and there are no further detailed insights/suggestions, which are not enough for AI computational models.

The documentation is a practical tool to provide clarifications for stakeholders in software engineering. Such documentation techniques have been adapted and developed to communicate key facts about machine learning systems. For example, FactSheets (Arnold et al. 2019) provide declarations of conformity for AI solutions with multi-dimensional details that capture various aspects of the AI model and its development processes to make it trustworthy by customers. Datasheets for Datasets (Gebru et al. 2021) aims to provide key information about the datasets used to develop machine learning models. Model Cards (Mitchell et al. 2019) are short documents that provide benchmarked evaluations in a variety of conditions for trained machine learning models. Model Cards also provide details of the model performance evaluation procedures, the context where models are to be used, and other relevant information. While the information provided by above methods is essential for stakeholders to evaluate whether an AI system meets their requirements, they are still less actionable. For example, machine learning developers need guidance to mitigate potential shortcomings to improve the system's performance while above methods may not provide such guidance. Method Cards (Adkins et al. 2022) aim to communicate key information with both prescriptive and descriptive elements about machine learning methods to guide machine learning developers throughout the process of model development to ensure that they are able to use those methods properly. Despite the usefulness of these approaches toward transparency and accountability of the machine learning system, they do not cover all aspects needed for the assurance of ethics and are limited from a user requirements' perspective on data and machine learning models.

Frameworks are also developed for the assurance of AI ethics. For example, Zicari et al. (2021) outline a process named *Z*-inspection to assess trustworthy AI. The process includes three phases of setup, assess, and resolve phases. The setup phase uses a catalogue of questions to define an understanding of the expectations of various stakeholders involved in the assessment. The assess phase analyses the usage scenarios, maps the ethical issues to AI ethical principles, and verifies AI ethical principles. *Z*-inspection uses the definition of trustworthy AI given by the high-level European Commission's expert group on AI. The term of trustworthiness is more on the machine learning model's performance (Ashmore et al. 2021) instead of ethics.

Standards play a key role in AI ethics operationalisation. There are different voluntary standardisation and certification initiatives in AI, notably from IEEE (IEEE Computer Society/Artificial Intelligence Standards Committee (C/AISC) 2019, 2020) and NIST (Schwartz et al. 2022). Australia published the Voluntary AI Safety Standard in August 2024, which gives practical guidance to Australian organisations on how to safely and responsibly use and innovate with AI (National Artificial Intelligence Centre 2024). The standards aim to ensure that the development and deployment of AI systems is safe and can be relied on. Furthermore, organisations and big companies have published different tools from the perspective of design of AI. For example, Microsoft[1] published a human-AI experiences (HAX) toolkit with a series of guidelines for the design of AI solutions. Google's People + AI[2] proposed design patterns and guidance for human-centred AI across the AI product development flow. Liao, Pribi´c, Han, Miller, and Sow (2021) from

[1] https://www.microsoft.com/en-us/haxtoolkit/

[2] https://pair.withgoogle.com/guidebook

IBM Research present a question-driven design process for explainable AI user experiences. The question-driven design process considers choices of explainable AI techniques, user needs, design, and evaluation of user experiences in the user questions. These works are important steps for the assurance of AI ethics. However, they still lack actionable tools for the AI ethics assurance or do not consider specific user requirements on ethics for targeted assurance.

2.2 Assurance case for mission critical systems

Machine learning is often a major component in cyber-physical systems (CPSs) in safety–critical domains (e.g. financial, medical, and automotive domains). The failure of machine learning–driven software in such systems can lead to the loss of human lives and/or spend significant costs to fix problems. The assurance of the safety for these systems is conducted from different perspectives: 1) Assessment of confidence and uncertainty (Duan et al. 2017). The confidence in an assurance case is defined as "the quality or state of being certain that the assurance case is appropriately and effectively structured, and correct" (Grigorova and Maibaum 2013). One approach to gain confidence in the assurance case is to analyse uncertainty, which can be dealt with through qualitative analysis or quantitative analysis. 2) Assessment of structure and content: software tools are developed to support assessors for processing assurance case documents such as structural checks and content checks: three categories of methods (correctness and completeness checks, structural constraints, user queries) are summarised for assessing an assurance case's structure, and five categories (argument assessment, evidence assessment, assessment interaction, assessment tracking, assessment reporting) are summarised for assessing an assurance case's content (Maksimov, Kokaly, and Chechik 2019). 3) Security assessment (Mohamad, Steghöfer, and Scandariato 2021): Security assurance cases are used to evaluate the security properties of a software system. It requires a thorough threat analysis throughout the software structure and at different stages of the software development. 4) Assurance weakeners (Shahandashti, Belle, Lethbridge, Odu, and Sivakumar 2023): The presence of assurance weakeners such as assurance deficits and logical fallacies in assurance cases indicates incomplete knowledge, evidence, or gaps in verification. These weakeners can affect user confidence in assurance arguments, interfering the successful assurance of the system. Shahandashti et al. (2023) conduct a systematic review and report a taxonomy that categorises assurance weakeners and related management approaches at the modelling level. This taxonomy can help practitioners to have better understanding of weakeners of existing assurance solutions. Based on the understanding, practitioners can devise more effective solutions to manage assurance weakeners.

Machine learning development is an iterative process. Therefore, an assurance case needs to continuously update related assurance artefacts such as dataset information and performance changes. Cârlan, Gauerhof, Gallina, and Burton (2022) examine how changes in the ML process affect safety arguments, with a focus on the adequacy of the data used to develop ML components. Yap (2021) presents a process-based framework for certifying trustworthy ML systems by adopting a simplified version of the common criteria used for certifying information technology and security software. Users can rely on the certification process-based framework to build confidence in the trustworthiness of an ML system. The certification is divided into two components: certifying basic system components which are covered by a disclosure such as FactSheets (Arnold et al. 2019) and certifying desired properties.

Picardi et al. (2020) introduce patterns for developing assurance arguments to demonstrate the safety of ML components. These argument patterns offer reusable templates for the necessary claims in an argument. They also develop a process which comprises five ML lifecycle phases (requirement elicitation, data management, model learning, model verification, and model deployment) and four-level assurance and development activities (assurance artefacts, assurance activities, development activities, and system artifacts), to establish the argument patterns and construct the assurance case for ML components. Assurance requirements for each ML lifecycle phase are derived based on a set of desiderata for each lifecycle phase (Ashmore et al. 2021). The patterns and process serve as a practical guide for the justifiable deployment of ML models in safety-related systems. These previous work for mission critical systems motivates us to develop AI assurance cases by considering user requirements and evidence in AI ethics.

3 Preliminaries

This section introduces concepts of assurances cases and system theoretic process.

Analysis that are used to set up AI ethics assurance approaches proposed in this paper.

3.1 Assurance cases

The term "assurance" is defined by the Cambridge Dictionary as "a promise to tell something to someone confidently or firmly, or a promise to cause someone to feel certain by removing doubt", and has long been used alongside objectives for safety, security, and other concerns (Rinehart et al. 2015). The term "assurance case", a generalisation of a

safety case, has become familiar to many practitioners (IEEE 2022). An assurance case (AC) is defined as "a collection of verifiable claims, arguments, and evidence assembled to confirm that a particular system/service aligns with given requirements" to solve the question of "are we sufficiently certain that this system is acceptably [safe, secure, reliable, etc.]?" (Mansourov and Campara 2010). An assurance case is usually represented either as plain text (unstructured natural language text, semi-structured text) or graphical notations that allows communication among various system stakeholders (e.g. suppliers, users, regulators). Many notations and tools have been proposed to create assurance arguments such as goal structuring notation (GSN) (Chelouati et al. 2023).

Furthermore, AI systems are sociotechnical, consisting of significant human and software elements. Goal modelling has proven useful in capturing stakeholder needs and documenting arguments and justifications for compliance, trade-offs, and decision-making in such systems (Horkoff et al. 2019). The user requirements notation (URN) (Dongmo and Van der Poll 2023; ITU-T Z.151 2018), which is based on goal modelling approach, is one of popular notations that are used in the safety assurance of mission critical systems (Sharifi et al. 2022). URN (Dongmo and Van der Poll 2023; ITU-T Z.151 2018) is a notation used for eliciting, analysing, specifying, and validating requirements. It enables software and requirements engineers to identify and define requirements for both proposed and evolving systems, as well as analyse them for correctness and completeness. It combines goal and process views used in requirements engineering activities, regulatory compliance, and regulatory intelligence (Sharifi et al. 2022). It is widely used in telecommunications systems, services, and business processes, as well as the high-level design of systems.

URN integrates two sub-languages: the goal-oriented requirement language (GRL), for modelling actors and their intentions, and use case maps (UCM), for describing scenarios and architectures. GRL supports goal-oriented modelling and reasoning, particularly for non-functional requirements and quality attributes. It features five main categories of concepts (see Fig. 1): elements, satisfaction levels, link composition, links and contribution types. UCM focuses on visually representing the causal flow of behaviour, optionally overlaid on a structure of components. It includes process-oriented concepts such as start and end points, responsibilities or activities, their corresponding system components or actors, and various relationships, including sequencing, guarded choices, concurrency, process decomposition, timers, and more (Amyot and Mussbacher 2011).

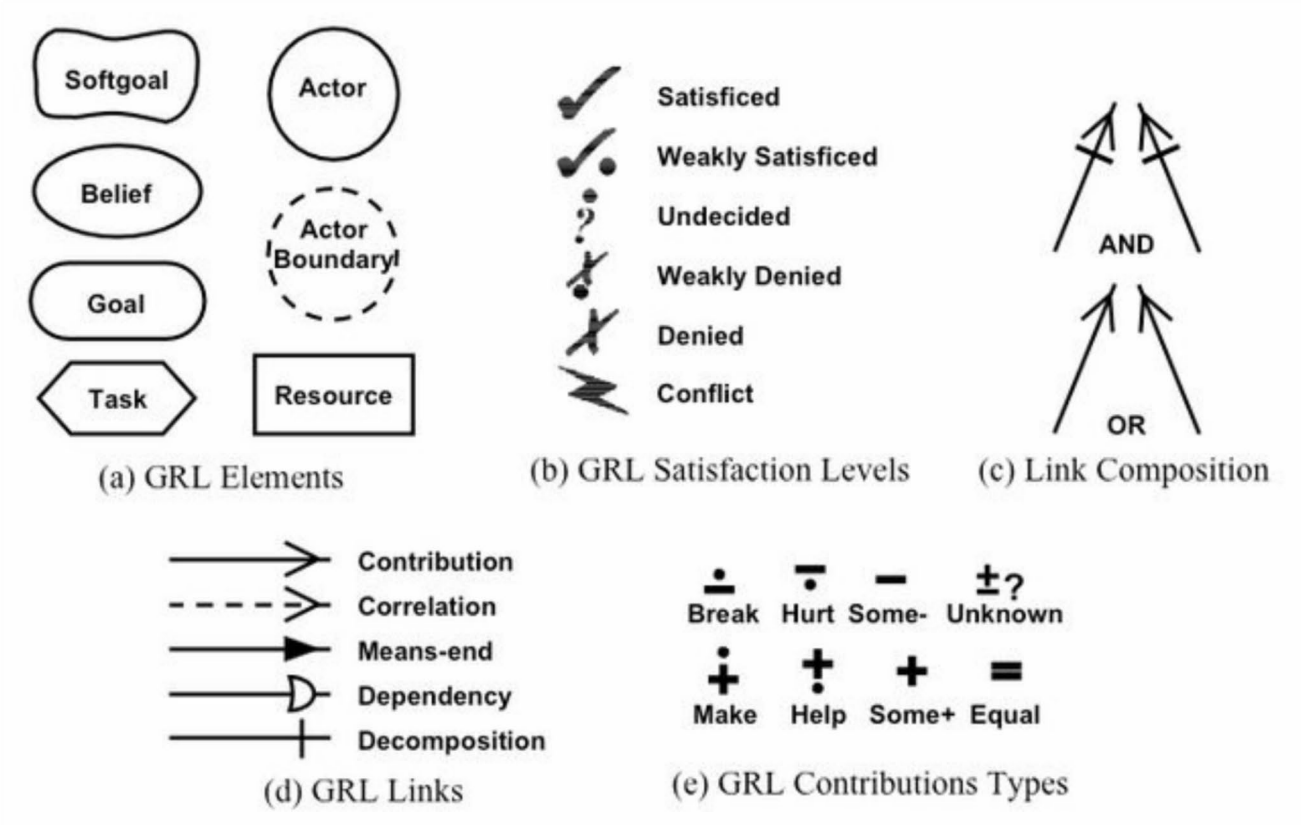


Fig. 1 Categories of concepts in goal-oriented requirement language (Weiss and Amyot 2007)

3.2 System theoretic process analysis

System theoretic process analysis (STPA) (Leveson 2016) is a hazard analysis method based on the system theoretic accident model and processes (STAMP) that evaluates potential systemic events where undesirable outcomes may occur. It offers a systematic approach to system safety and ensures clear traceability between hazards and design. Figure 2 illustrates the STPA pipeline. In this pipeline, STPA defines three system states: 1) loss states, which are unacceptable to stakeholders (e.g., loss of fairness, assets, or reputation); 2) hazard states, where no loss has occurred yet but which could lead to a loss state; and 3) safe states, which are acceptable to stakeholders. STPA uses these concepts within a hierarchical control structure to identify potential pathways through which the system might transition to a hazard state. As shown in Fig. 2, the control actions that lead to hazards referred to as unsafe control actions (UCAs). The UCAs are used as a basis to identify causal scenarios that might cause the UCAs. Safety constraints are derived from causal scenarios, which are then used to define design recommendations to mitigate the hazards (Sharifi et al. 2022). System safety requirements are identified for the overall system safety assurance.

STPA offers a broad definition of loss that can be adapted to different types of losses in other safety–critical areas. Additionally, STPA provides a methodology for systematically defining system controls that are traceable to the losses stakeholders aim to prevent (Sharifi et al. 2022). Such features much benefit a requirement-based certification approach.

4 User requirement–oriented AI ethics assurance

This section first introduces the concept of AI ethics assurance case. A user requirement–oriented approach for AI ethics assurance framework is then proposed. The detailed AI ethics assurance process is also presented in this section. The traceable link between user requirements on ethics and AI ethics assurance is the focus of this section.

4.1 AI ethics assurance case

Assuring AI ethics requires more than simple reports from validators. A rigorous and systematic approach to capturing and communicating the rationale throughout the AI lifecycle is essential for certifying AI systems. In this paper, the concept of assurance case is adapted into AI ethics for the AI ethics assurance. An AI ethics assurance case is a comprehensive, defensible, and valid justification of the ethical considerations for a specific AI system and its application. The assurance case therefore provides the grounds for user confidence and trust in AI solutions from the ethical perspective. Similar to system safety assurances, the AI ethics assurance case acts as the basis for the decision making in certification of AI solutions.

Adapted from the safety assurance cases (Picardi et al. 2020), the AI ethics assurance claim is that the AI model meets the defined ethical requirements, including ethical principles relevant to the specified operating environment. The claim focuses on the specific AI ethical principles to reflect the AI model's impact on the overall system's ethics. Furthermore, the verification evidence that supports the claim must be specific evidence that shows the AI ethical principles are satisfied. The assurance argument for an AI model should be viewed as a component of a broader assurance argument addressing the safety of the entire system. AI ethics assurance involves multiple concerns (business, regulations, AI, data) and must clearly convey compliance and risk assessments to certification authorities.

4.2 AI Ethics assurance framework—a user requirement–oriented approach

Motivated by the system theoretic process analysis in safety critical systems (Sharifi et al. 2022), this paper proposes a requirement-oriented approach for AI ethics assurance cases, focusing on addressing ethical requirements rather than merely arguing that AI solutions are free

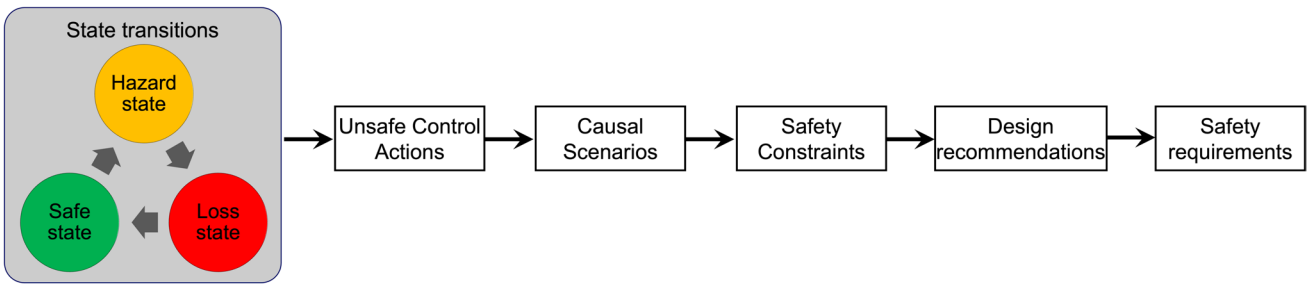


Fig. 2 System theoretic process analysis pipeline

from ethical hazards or risks. Figure 3 shows the ethical process analysis from ethical requirements' perspective. In this process, three states of the AI ethical status are defined: 1) ethical safe state or ethical state which is the safe state in ethics; 2) ethical loss state or loss state that is unacceptable to the stakeholders of the AI system from the ethical perspective, e.g. leak of privacy, lack of transparency; and 3) ethical hazard state or hazard state, where no ethical loss has occurred yet but which could transition to an ethical loss state. Similar to STPA for safety assurance cases (Sharifi et al. 2022), these states can be transitioned from one state to another. A hierarchical control structure is used for such state transitions. This ethical process analysis is divided into two processes (see Fig. 3): derive design recommendations from state transitions and assure AI ethics based on ethical requirements. In the process of deriving design recommendations from state transitions, the unethical AI actions (UAIAs) that can lead to ethical hazards are firstly identified. The UAIAs are used as a basis to identify ethical causal scenarios that might cause the UAIAs. Ethical constraints are then derived from causal scenarios. Design recommendations are therefore introduced based on constraints. Finally, ethical requirements are identified to mitigate the ethical hazards in the state transitions. On the other side, the ethical requirements are used to assure AI ethics by checking components from ethical requirements to UAIAs in the assurance case (see Fig. 3).

Besides requirements, this paper also introduces evidence and validation into the AI ethics assurance framework. These three components of requirements, evidence, and validation (REV) are called three pillars of AI ethics assurance cases (see Fig. 4). User requirements on AI ethics are user's specific interest in AI ethics and they act as the basis for the assurance. The evidence is any supporting information that is used to justify the ethical trustworthiness of AI systems for their intended applications. The validation is the step that is conducted by experts in AI ethics and/or regulators so that AI systems can be endorsed and used ethically. Based on these three pillars,

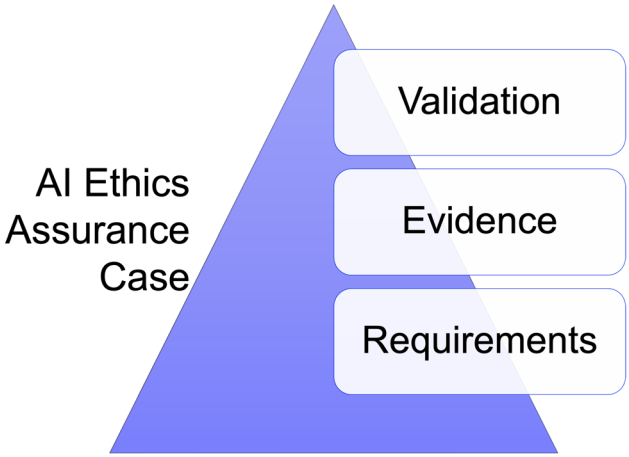


Fig. 4 Three pillars in AI ethics assurance cases

the AI ethics assurance case is set up using system theoretic process analysis.

Based on the ethical process analysis, URN is used to model the AI ethics assurance process in this paper. This process includes the following steps:

1. Identify stakeholders.
2. Develop a strategic dependency model.
3. Detail the functional goals.
4. Conduct STPA.
5. Create assurance cases.

The details of these steps are described in the following subsections.

4.3 Identify stakeholders

The first step in the AI ethics assurance process is to develop a comprehensive understanding of AI solution stakeholders and their dependencies, known as a strategic dependency model in goal-oriented requirements engineering (Sharifi

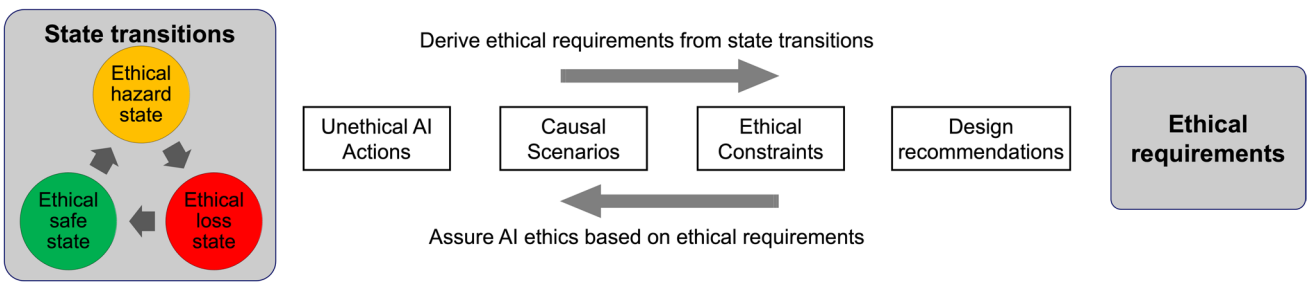


Fig. 3 Ethical process analysis from ethical requirements' perspective

et al. 2022). Seven groups of stakeholders are identified in the AI ethics assurance cases:

- AI suppliers: Those responsible for creating, providing, and supporting the AI system in the assurance
- AI supplier administrators: Those who have administrative responsibility for the AI system.
- AI regulators: Those who are responsible for overseeing and enforcing regulations related to the development, deployment, and use of AI based on the assurance case for the AI system.
- System administrators: Those responsible for administrative role on the assurance system (e.g. E-LENS in our study).
- Ethics validators: Those who check evidence provided by AI suppliers for the compliance of the AI system on AI ethical principles.
- AI users: Those who rely on AI solutions to inform their decisions.
- Visitors/general public: The public is given access to the assurance system and the assurance case to understand its function, ensure transparency, and assess any residual risks.

Among these stakeholders, AI supplier administrators, system administrators, and visitors/general public are the administrators of assurance systems or general public, and they could not be considered in the strategic dependency model. Figure 5 shows the strategic dependency model in the proposed AI ethics assurance case based on the identified stakeholders. It allows for the identification of high-level goals of various stakeholders and their inter-dependencies.

4.4 Detail the functional goals

The high-level goals identified in the strategic dependency model can be further refined with the use of UCM described in URN. UCM gives a visual scenario of functional goals of the system and actors' behaviour. This section uses the AI ethical principle of transparency as an example to show how URN is used for the AI ethics assurance case. Figure 6 shows the UCM for the transparency in the AI ethics assurance case. In this UCM, transparency is assured by checking the traceability, communication, and explainability sequentially. The failure of the assurance of any component results in the failure of the assurance of the transparency.

4.5 Conduct STPA

With the top-level goals identified in the previous step using UCM, the ethical loss states and their corresponding ethical hazards are defined to establish the ethical boundaries of the AI solution. Table 1 shows the list of example ethical losses for transparency, and Table 2 presents the list of example ethical hazards corresponding to ethical losses in transparency in Table 1.

With the definition of ethical losses and hazards, it is necessary to evaluate possible ways in which any ethical losses and hazards can occur. According to the theory of System theoretic accident model (STAM) and processes

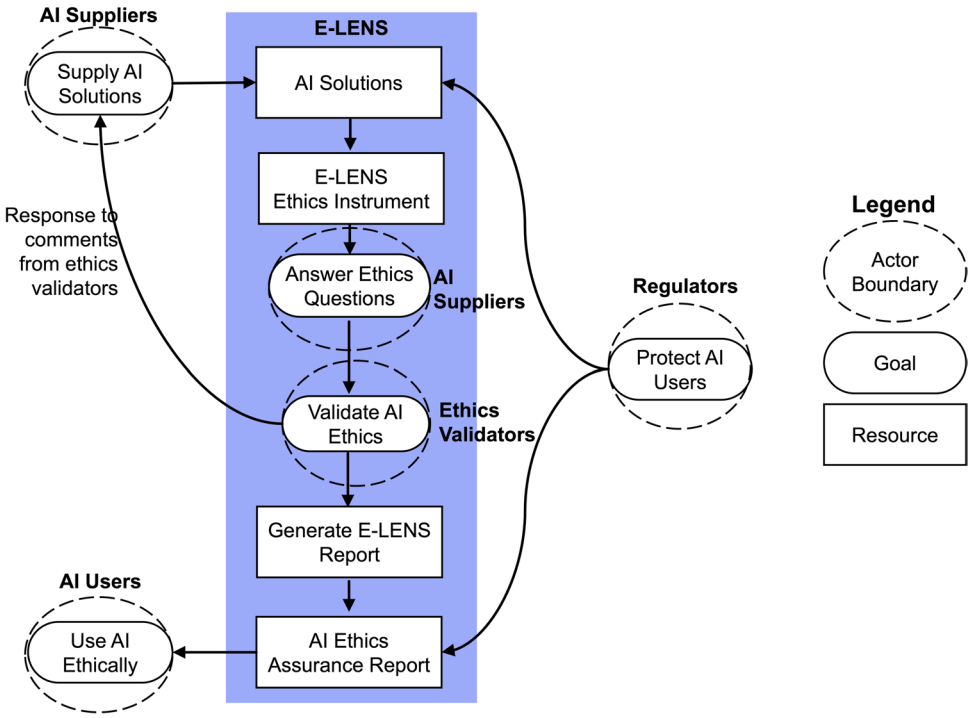


Fig. 5 Strategic dependency model in the AI ethics assurance case

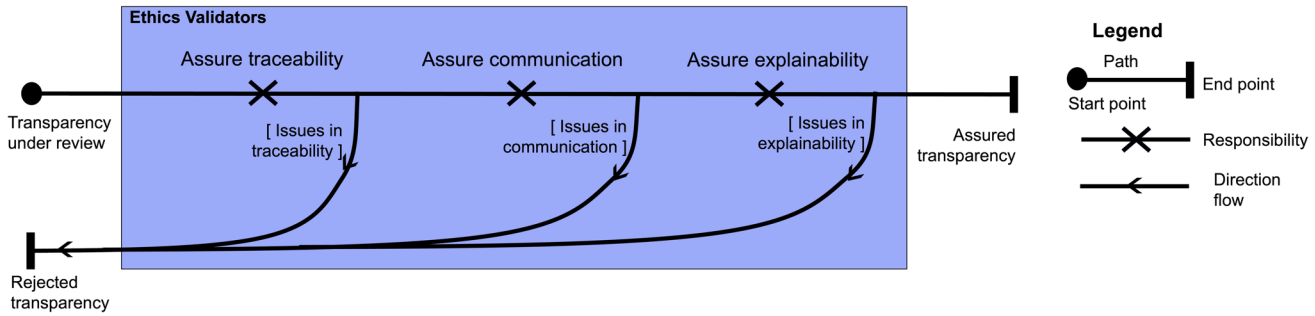


Fig. 6 Use case map for the transparency

Table 1 List of ethical losses in transparency

Principle segment	ID	Ethical losses
Traceability	L1	Loss of documentation related to data, its collection, and preprocessing
	L2	Loss of the documentation of the information on the method of training and testing the algorithm
	L3	Loss of documentation regarding the data used for testing and validation
	L4	Loss of the documentation of the outcomes of or decisions taken by the algorithm
Communication	L5	Loss of clear communication about the AI system's characteristics, limitations, and risks
Explainability	L6	Loss of the ability to provide an explanation for why the system made a particular choice leading to a specific outcome, in a way that is understandable to all users
	L7	Loss of the evaluation of the faithfulness of explanation, i.e. are relevance scores in explanations indicative of "true" importance
	L8	Loss of evaluating the monotonicity of the explanation, meaning whether adding more positive evidence increases the likelihood of classification in the specified class

Table 2 List of ethical hazards in transparency

Principle segment	ID	Ethical losses	Links to losses
Traceability	H1	Required documentation of the information on the data, collection, and preprocessing not obtained	L1
	H2	Required documentation of the information on the method of training and testing the algorithm not obtained	L2
	H3	Required documentation of the information on the data used to test and validate not obtained	L3
	H4	Required documentation of the algorithm's outcomes or decisions was not obtained	L4
Communication	H5	There is no communication regarding the AI system's characteristics, limitations, or risks	L5
Explainability	H6	Explanations on the model or outputs not provided	L6
	H7	Explanations on the model or outputs are wrong	L6
	H8	Evaluation of the faithfulness of explanation not provided	L7
	H9	The faithfulness of explanation is low	L7
	H10	Evaluation of the monotonicity of explanation not provided	L8
	H11	The monotonicity of explanation is low	L8

used in STPA (Sharifi et al. 2022): STAM treats systemic qualities as a dynamic control issue. A control diagram would help to visually understand the relations of different components and how they affect each other. Figure 7 shows the control diagram regarding transparency for an AI model building in the AI lifecycle. In this diagram, the green-shaded boxes are related to the ethical principle segment of traceability, the purple-shaded boxes are related to the ethical principle segment of communication, and the orange-shaded boxes are related to the ethical principle segment of explanation. From this control diagram, we can easily see how different actions affect the transparency of which component in the AI lifecycle, for example "document the training method" affects the traceability of "model building" and "explanation to model/outputs" (action A) affects the explainability to "AI users".

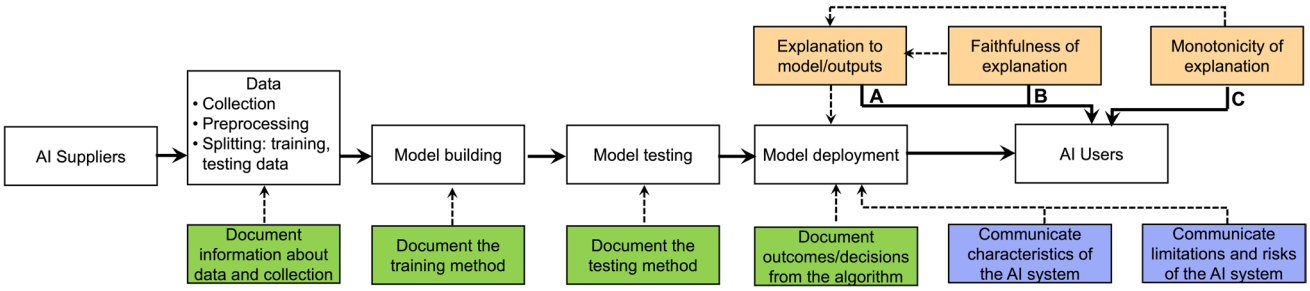


Fig. 7 Control diagram regarding transparency for an AI model building in the AI lifecycle

Table 3 Samples of unethical AI actions (UAIAs) identified for the control actions A, B, and C

Control action	Provided	Not provided
A:Explain model/outputs	UAIA101: The AI system provides the wrong explanation [H7]	UAIA104: The AI system does not explain the model/outputs [H6]
B: Evaluate the faithfulnessof explanations	UAIA102: Users are not satisfied with the faithfulness of the explanation. [H9]	UAIA105: The AI system does not evaluate the faithfulness of explanations. [H8]
C: Evaluate the monotonicity of explanations	UAIA103: Users are not satisfied with the monotonicity of the explanation. [H11]	UAIA106: The AI system does not evaluate the monotonicity of explanations. [H10]

Table 4 List of causal scenarios for UAIA101 (the AI system provides the wrong explanation)

ID	Causal scenarios
CS1	The AI system used the wrong explanation method
CS2	The explanation method used the wrong parameters
CS3	The data used to generate explanations was not appropriate
CS4	The system lacked documentation on the explanation method

Taking the control action of A (provide "explanation to model/outputs") in Fig. 7 as an example, it is necessary to identify unethical AI actions for the action A based on the proposed framework (see Fig. 3). Table 3 illustrates examples of UAIAs identified for the actions of A, B, and C (actions can be hazards) when those actions are provided or not provided.

After UAIAs are identified, the causal scenarios for UAIAs need to be developed based on UAIAs according to the proposed framework (see Fig. 3). Table 4 provides examples of causal scenarios for UAIA101 (the AI system provides the wrong explanation) identified in the previous step.

As shown in Fig. 3, the developed causal scenarios for unethical AI actions are used to elicit ethical constraints and design recommendations. Table 5 lists ethical constraints related to the hazards for transparency. Table 6 shows examples of design recommendations for UAIA101 for the ethical constraint of EC101 (the AI system must not provide the wrong explanation). As profiled in Fig. 3, the design recommendations are then further refined into ethical requirements for AI solutions. Table 7 lists examples of ethical requirements elicited from design recommendations for UAIA101. These ethical requirements are used for the AI ethics assurance case. Table 7 also provides examples of verification methods for ethical requirements, such as demonstration, black-box testing, scenario testing, and algorithmic evaluation.

Based on these previous steps from different ethical state identification to the identification of ethical requirements for an AI solution, a URN-based assurance case for the AI ethics is created. Figure 8 shows an example of assurance case for the explainability. This assurance case focuses on the ethical risks that the AI solution may expose through the ethical requirements. From GRL representations of Fig. 8, users can visually assess the ethical issues following the graph and confirm that "all ethical hazards are mitigated". Rather than arguing that the system is free from ethical hazards, as many previous efforts have done, the focus is on addressing user requirements related to ethical risks and providing evidence to mitigate those risks.

5 E-LENS: an extensible platform for AI ethics assurance cases

In order to perform the AI ethics assurance case as shown in Fig. 8 easily by users, this paper proposes a platform named E-LENS (Ethical-Lens) and incorporates AI ethical requirements into this platform to implement the user requirement–oriented AI ethics assurance. This platform is designed to be extensible so that new components, such as

Table 5 List of ethical constraints related to the ethical hazards for transparency

Principle segment	ID	Hazard Ethical constraints
Traceability	H1	EC1: The information on the data, data collection, and data preprocessing should be documented.
		EC2: If the system loses this data documentation, it should be re-documented.
	H2	EC3: The information on the method of training and testing the algorithm should be documented.
		EC4: If the system loses this algorithm documentation, it should be re-documented.
	H3	EC5: Documentation of the data used to test and validate the model is required.
		EC6: If the system loses this test data documentation, it should be re-documented.
	H4	EC7: The outcomes or decisions made by the algorithm must be documented.
		EC8: If the system loses this outcome/decision documentation, it should be re-documented.
Communication	H5	EC9: The characteristics, limitations, and risks of the AI system should be communicated to users.
		EC10: If the communication of characteristics, limitations, and risks of the AI system to end users is lost, it should be re-established.
Explainability	H6	EC11: Explanations on the model or outputs should be provided.
	H7	EC12: If explanations on the model or outputs are wrong, the explanations should be re-produced.
	H8	EC13: The faithfulness of explanation should be provided.
	H9	EC14: If users are not satisfied with the faithfulness of explanation, the explanation should be re-produced.
	H10	EC15: The monotonicity of explanation should be provided.
	H11	EC16: If users are not satisfied with the monotonicity of explanation, the explanation should be re-produced.

Table 6 List of ethical constraints and design recommendation for UAIA101 (The AI system provides the wrong explanation)

ID	Ethical constraint/Design recommendation
EC101	The AI system must not provide the wrong explanation.
DR101.1	System should require the confirmation that there are explanations provided.
DR101.2	System should provide the faithfulness of explanations.
DR101.3	The evaluation of faithfulness of explanations should use data with distributions similar to the distributions of actual input data.
DR101.4	System should provide documents on the method of explanation and method that is used to evaluate faithfulness of explanations.

Table 7 List of ethical requirements elicited from design recommendations for UAIA101 (The AI system provides the wrong explanation)

ID	Ethical requirement	Traceability to DRs	Verification method
R101.1	System should display explanations besides the AI outcomes/decisions.	DR101.1	Demonstration
R101.2	System should display the faithfulness of explanations.	DR101.2	Algorithmic evaluation
R101.3	System should confirm that the distribution of data used for the evaluation of faithfulness of explanations is similar to distributions of actual input data.	DR101.3	Black-box testing
R101.4	System should document the method for the explanation and evaluation of faithfulness of explanations.	DR101.4	Demonstration

new ethical principles and ethical questions related to ethical requirements, can be added if necessary.

E-LENS has the following modules including verification methods and other features (see Fig. 9):

- Checklist instrument: A checklist of questionnaire, which corresponds to ethical requirements derived in the assurance case, helps users to easily check each ethical requirement, and provides evidence for the assurance.
- Qualitative assessment: Considering the nature of ethics, qualitative approaches are used to assess ethical requirements.
- Algorithmic assessment: Given the data-driven nature of machine learning and AI, algorithmic approaches are used to answer ethical requirements in the assurance.
- Iterative feedback mechanism: The AI ethics assurance is designed as an iterative process. AI solution providers and ethics validators communicate in an iterative

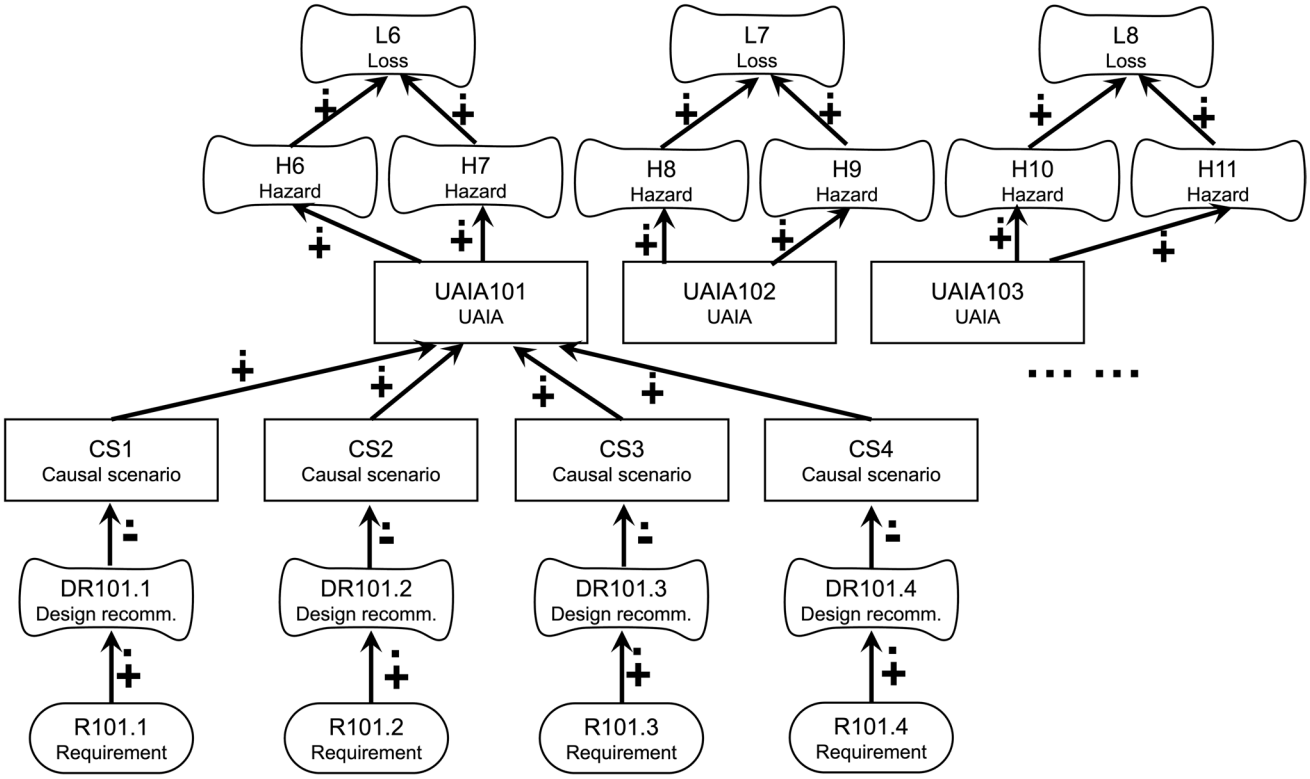


Fig. 8 The AI ethics assurance case for the explainability

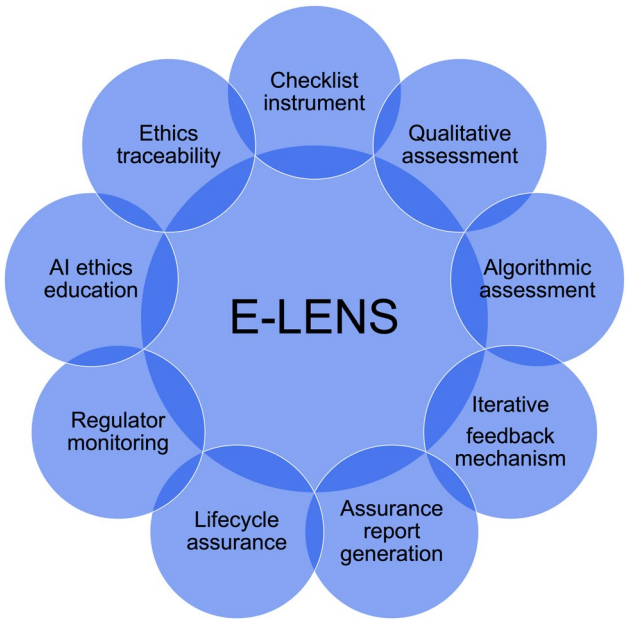


Fig. 9 Modules included in the E-LENS framework

feedback loop to make sure all ethics requirements are met.

- Assurance report generation: The assurance report documents the overall AI ethical requirements and their evidence or answers from AI solution providers, which allows the easy communication among stakeholders regarding AI ethics.
- Lifecycle assurance: The assurance encompasses the entire AI lifecycle, ensuring ethical safety at every stage.
- Regulator monitoring: E-LENS introduces the regulator into the AI ethics assurance pipeline and allows the AI ethics monitoring in real-time by regulators.
- AI ethics education: Education on AI ethics helps stakeholders better understand ethical principles, leading to more effective implementation of those principles in AI systems.
- Ethics traceability: The evidence for each AI ethics argument at every AI lifecycle stage is recorded in E-LENS, which allows to access and trace AI ethics status at any time.

Each of these features are introduced in more details in the following subsections.

Figure 10 shows the pipeline of the AI ethics assurance in E-LENS.

5.1 Checklist instrument

Despite the theoretical weaknesses in its framework, the checklist-style questionnaires are widely employed as

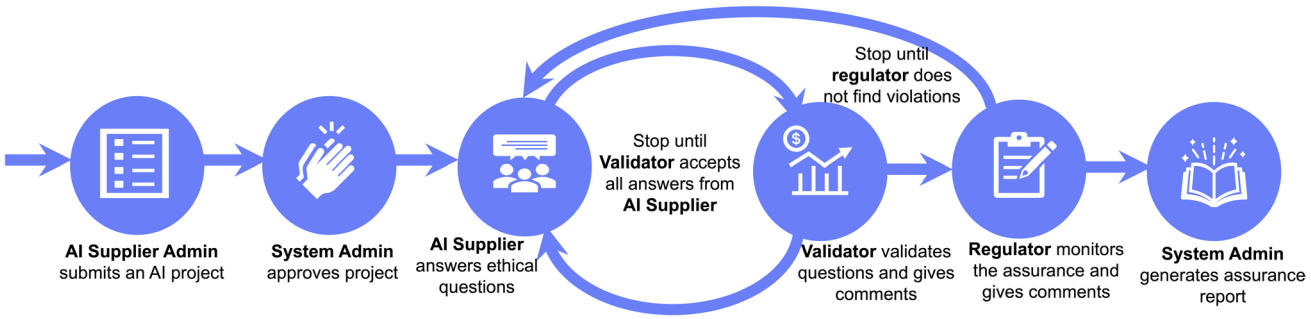


Fig. 10 The pipeline of the AI ethics assurance in E-LENS

the standard ethics governance model in biomedical and clinical research (Canca 2019). Similar to checklist-style questionnaires employed in the biomedical and clinical research, our work proposes using a checklist as a tool to validate the ethics surrounding AI solutions. It is ideal to develop checklist-style questionnaires for each identified ethical principle to audit AI compliance with these principles.

Inspired by the assessment list for trustworthy AI published by the HLEG-AI of the European Commission (Ala-Pietilä et al. 2020), this paper develops a hierarchical checklist questionnaire approach for ethical requirements to assure AI ethical principles. Each ethical principle has an overarching question on ethical requirements which is split into different sub-questions from different perspectives (segments). For example, the transparency principle is evaluated from the perspectives (segments) of requirements on traceability, communication, and explainability (see Fig. 11). The previous survey indicates that ethical principles for AI commonly converge on transparency, justice and fairness, responsibility and accountability, non-maleficence, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity (Zhou et al. 2020). This paper focuses on mandatory ethical principles of transparency, fairness, accountability, and privacy. We categorise the existing ethical principle questions into different requirements for different AI lifecycle stages and desiderata (Ashmore et al. 2021): relevant, complete, balanced, and accurate. Furthermore, the hierarchical checklist approach is designed to be extensible in E-LENS, which allows users to add new ethical principles, new principle segments, and new checklist questions if necessary in the E-LENS framework.

5.2 Qualitative assessment

Ethics and AI are two distinct scientific domains. Ethics is a highly abstract social science, whereas AI involves defining strict quantitative metrics, such as prediction accuracy. Therefore, we propose implementing AI ethical principles both qualitatively and quantitatively/algorithmically.

The qualitative assessment is designed as checklist questions that are answered with qualitative descriptions. In the E-LENS framework, qualitative questions are classified into two types: multiple choice questions and extended response questions. Multiple choice questions ask users to select one answer from provided options, while extended response questions require users to write short descriptions as answers on how AI solutions solve the ethical issues. For example, extended response questions are used to validate the measures implemented for protecting data privacy.

5.3 Algorithmic assessment

Quantitative or algorithmic assessment aims to evaluate how well AI solutions adhere to ethical principles by developing algorithmic approaches for this purpose. For instance, quantitative measures can be used to validate the explainability of AI solutions and assess the fairness of both data and models. Various approaches can be used to enhance the transparency of an AI model or algorithm (Zhou et al. 2021), assess algorithmic

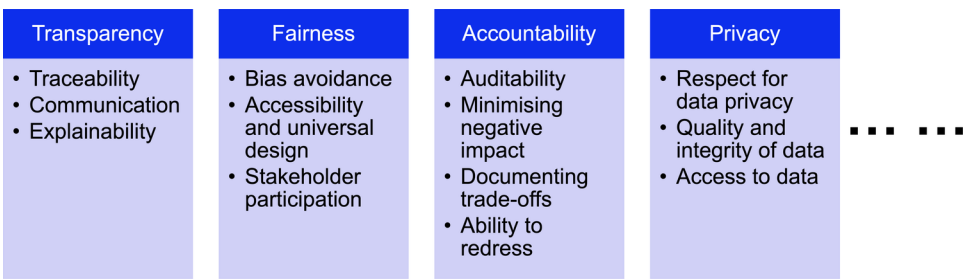


Fig. 11 Ethical principles and their segments

bias for fairness in predictions (Bellamy et al. 2018), or improve privacy through algorithms (Kairouz et al. 2021). These algorithmic methods can be effective for AI experts and developers in detecting and assessing ethical issues within AI models.

In the E-LENS framework, algorithms are developed to evaluate the AI model explainability as well as fairness of data and AI models. They require users to provide relevant data and model information for the evaluation.

5.4 Feedback mechanism

The AI ethics assurance process is designed as an iterative looping process, where AI suppliers answer ethical questions and the answers are then evaluated by AI ethics validators (see Fig. 10). AI ethics validators give feedback to AI suppliers for any further improvements until AI ethics validators accept all answers from AI suppliers.

5.5 Assurance report generation

AI ethics assurance report documents the ethics status and support information on AI solutions. It helps to communicate among stakeholders on AI ethics. In the E-LENS framework, all answers to ethics questions are saved in a database. A report generation function is developed to create PDF documents to stakeholders on the AI ethics status and support information. Two types of report are provided in the E-LENS framework: 1) the full report which documents all details of answers including comments from AI ethics validators to all AI ethics questions; and 2) the summary report which provides short summary for the ethics status of each ethical principles.

5.6 Lifecycle assurance

The lifecycle of a typical AI application typically encompasses several stages: business and use-case development, design, data collection for training and testing, model building and testing, model deployment, and system monitoring (see Fig. 12). This framework provides a high-level view of how an AI project should be structured to deliver practical business value at each stage. Morley et al. (2020) emphasise that ethical principles should be integrated at every stage of the AI lifecycle to ensure that the AI system is designed, implemented, and deployed ethically.

In the E-LENS framework, checklist questions are designed for every ethical principle at each stage to make sure that the AI ethics is assured covering the AI lifecycle.

5.7 Regulator monitoring

Although it is not fully established in many countries, the regulation framework for AI ethics plays an important role in the success of assurance of AI ethics. In the E-LENS framework, regulators are introduced into the assurance pipeline of AI ethics as shown in Fig. 10. In this pipeline, after the validator finishes the assessment, the regulator checks the assessment results to monitor whether there are any violations to related regulations. If yes, the assurance process returns to the previous step and the AI supplier is required to fix any comments from the regulator for further assessment until the regulator does not find further violations to related regulations (see Fig. 10).

5.8 Ethics traceability

Ethics traceability refers to the capability that relevant stakeholders can trace back to the previous ethics status or support evidence for ethical principles. It allows stakeholders easily learn ethical issues and make informed decisions. In the E-LENS framework, all of the ethics status and support evidence are saved in databases, stakeholders can easily access the information to trace back the assurance of AI ethics for AI solutions.

6 AI ethics assurance best practice: AI for recruitment

This paper presents the E-LENS framework for assuring ethical principles in AI, implemented as a web-based application to effectively ensure ethical compliance in AI

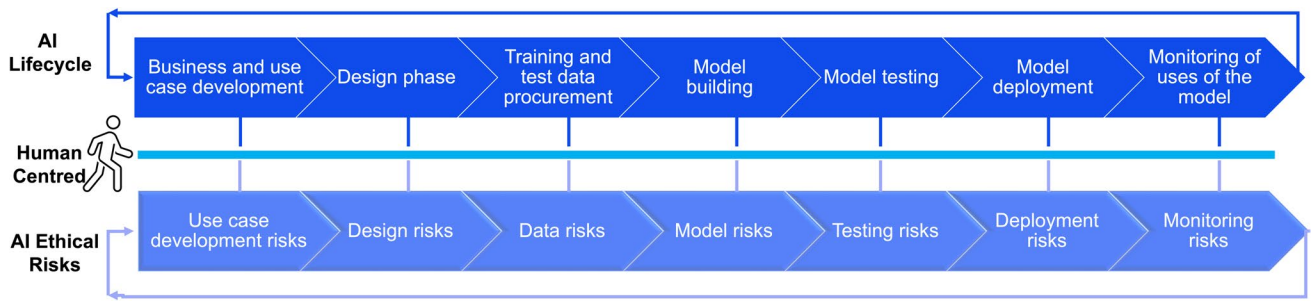


Fig. 12 The AI lifecycle and ethical risks

solutions. Designed to accommodate the evolving nature of ethical investigations, the platform is open and flexible, allowing for the addition or customisation of ethical principles and corresponding qualitative and quantitative questions to suit various project requirements. Upon completion of the validation process, the platform generates a summary report for AI solution providers, detailing the ethical aspects addressed and identifying areas for improvement from an ethical standpoint. In this section, we show a case study of the use of E-LENS in the AI ethics assurance.

In the human resources (HR) recruiting process, one of the most challenge steps is to shortlist a number of most qualified applications for a job position for interviews by reviewing each application (Gusdorf 2008). A large number of applications (even hundreds or thousands) are submitted if the posted position is popular. It is much time consuming if all applications are reviewed by humans one by one. Even more, different reviewers may have different comments on the same application, which cause further bias in the selection. Each application may include resume, cover letter, answers to criteria, certificates, and other supporting documents. HR departments often have datasets that include potential job candidates, past applicants, and current employees. AI is used to generate analytics-backed insights into various HR processes, including the shortlisting of candidates during recruitment (Kaushal et al. 2023; Roopalatha and Sucharita 2024).

However, AI may raise myriad ethical issues because of the use of human-related data. Some of typical ethical issues include transparency, data privacy, gender bias, and/or unclear accountability (Ajonbadi, Khan, and Owolewa 2024). The assessment of the AI algorithm's adherence to core ethical principles is therefore critical to delivering a responsible solution which proactively addresses potential risk.

In this section, the proposed user requirement–based E-LENS platform is applied to assure AI ethics of an AI-driven HR shortlisting system. Firstly, user requirements are defined from the ethical perspective for the shortlisting system such as free of bias to a certain gender and transparent in the shortlisting decision making. These ethical requirements act as the basis for the setup of the ethics assurance case. Following this step, a series of checklist-style questionnaires are used to gather evidence for assessing the ethics of AI solutions both qualitatively and quantitatively. For instance, qualitative questions validate measures used to protect data privacy, while quantitative questions evaluate AI compliance with ethical principles through specific metrics. Examples of quantitative

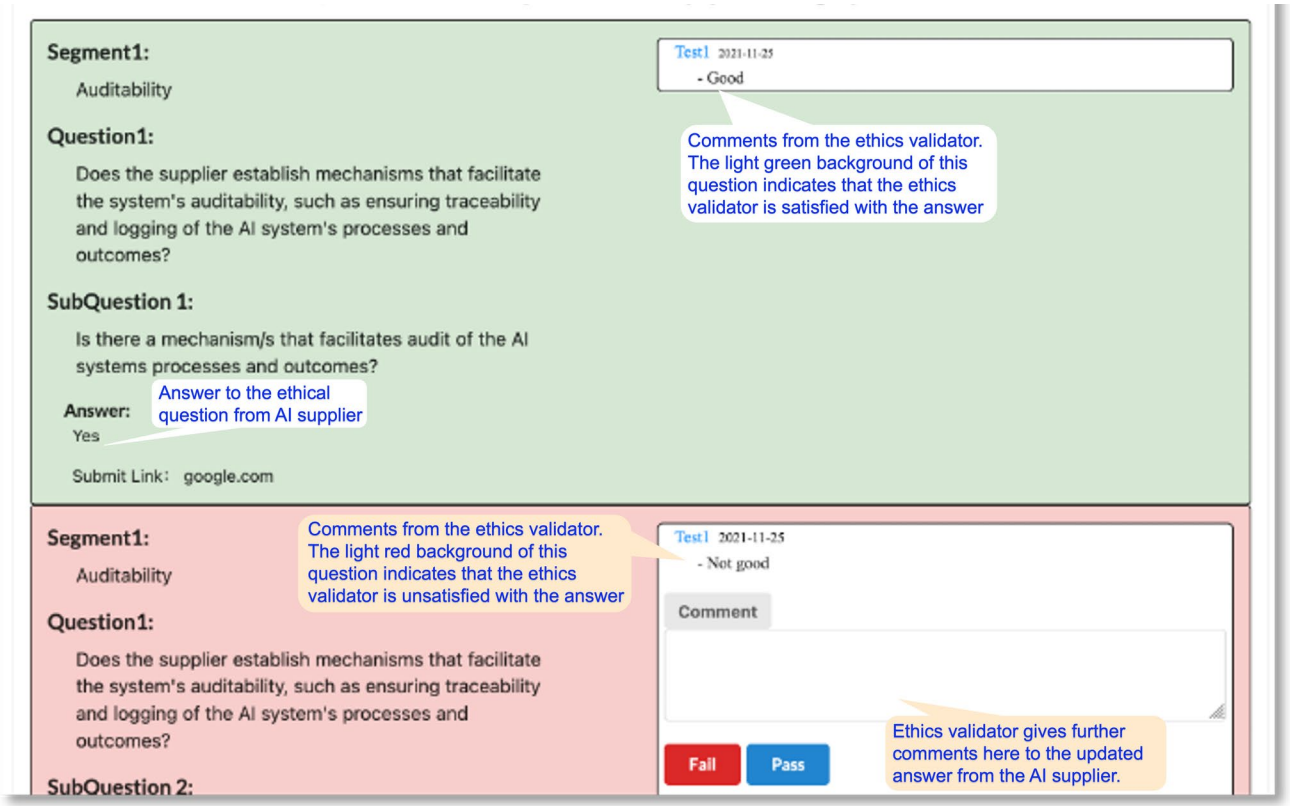


Fig. 13 The feedback mechanism implemented in E-LENS

measures include assessing the faithfulness (the relevance of features) and monotonicity (the effect of individual features on model performance) of AI explainability, as well as examining the fairness of both data and AI models. Ethics validators and regulators then check the answers to ethical questions to assure answers. As shown in Fig. 10, this is a looping process until validators and regulators are satisfied with the evidence as answers to questions. Figure 13 shows such feedback-looping mechanism implemented in E-LENS. In the E-LENS implementation, the ethics validator/regulator gives comments to the answer to the ethics question from the AI supplier. If he/she is satisfied with the answer, the background of this question will become light green and no further comments can be given after he/she submits comments. Otherwise, the background of this question will be light red and the ethics validator/regulator is allowed to given further comments to further answers until he/she is satisfied with the answer. Finally, the AI ethics assurance report is generated as a PDF file based on answers to ethical questions which are relevant to user ethical requirements.

7 Discussion and limitations

In this paper, three pillars of ethical requirements, evidence, and validation play the key role in an AI ethics assurance case. Ethical requirements are user's specific interest in AI ethics and they act as the basis for the assurance. The assurance case needs to provide "evidence" to support and justify the ethical trustworthiness of AI systems. With the support evidence, experts in AI ethics and/or regulators need to conduct "validation" so that AI systems can be endorsed and used ethically. However, this paper does not touch the details on how to identify ethical requirements especially for different domains. Furthermore, as reviewed in Sect. 2, AI standard plays a key role in AI ethics operationalisation. E-LENS does not consider the AI standard in the pipeline. We argue that the inclusion of AI standard in E-LENS will conduct more objective and trustworthy assurances. However, the AI standard itself is mostly under development. Besides, the integration of AI act into E-LENS will allow the comprehensive AI ethics assurance, AI ethics governance, and certification. These limitations will also be the future work of this paper.

8 Conclusion and future work

This paper introduces the concept of assurance cases, which is widely used in safety–critical systems for the safety assurance, into the AI ethics assurance for a new approach of user requirement–oriented AI ethics assurance. The new approach addresses the significant challenge of assurance of AI ethics by integrating user requirements, evidence, and validation into the AI ethics assurance case. This paper provides a robust solution that uses the widely accepted safety assurance theory in solving AI ethics assurances. The implementation of E-LENS for the user requirement–oriented AI ethics assurance cases allows the versatile and flexible assurance of AI ethics in practices. The future work of this paper will focus on the development of AI ethics assurance cases based on the proposed user requirement–oriented approach for specific domains or applications and integrate them into the E-LENS platform. Ideally, assurance case development occurs in parallel to the AI solution development, resulting in a legitimately strong argument for the desired ethical requirement.

Acknowledgements All authors thank reviewers' insightful comments to this paper.

Author contribution All authors have contributed to the formation, discussion, and writing of this article.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability Data are contained within the article.

Declarations

Ethical approval and consent to participate Not applicable.

Consent for publication Authors got consent for publication.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adkins D, Alsallakh B, Cheema A, Kokhlikyan N, McReynolds E, Mishra P, ..., Zvyagina P (2022) Method cards for prescriptive machine-learning transparency. Proceedings of the 1st international conference on ai engineering: Software engineering for AI pp 90–100 https://doi.org/10.1145/3522664.3528600

Ajonbadi HA, Khan S, Owolewa M (2024) Before it goes south: the ethical dilemma of artificial intelligence in human resource management—the bangladesh workplace experience. In: Adekoya OD, Mordi C, Ajonbadi HA (eds) HRM, artificial intelligence

and the future of work: insights from the global south. Palgrave Macmillan, Cham, Switzerland, pp 147–167. https://doi.org/10.1007/978-3-031-62369-1_8

Ala-Pietil¨a P, Bonnet Y, Bergmann U, Bielikova M, Bonefeld-Dahl C, Bauer W, ..., others (2020) The assessment list for trustworthy artificial intelligence (altai). Brussels, Belgium: European Commission

Amyot D, Mussbacher G (2011) User requirements notation: the first ten years, the next ten years. J Softw 6(5):747–768

Arnold M, Bellamy RK, Hind M, Houde S, Mehta S, Mojsilovi´c A et al (2019) Factsheets: increasing trust in ai services through supplier's declarations of conformity. IBM J Res Dev 63(4/5):6–1

Ashmore R, Calinescu R, Paterson C (2021) Assuring the machine learning lifecycle: desiderata, methods, and challenges. ACM Computing Surveys (CSUR) 54(5):1–39

B´elisle-Pipon J-C, Monteferrante E, Roy M-C, Couture V (2023) Artificial intelligence ethics has a black box problem. AI & Soc 38(4):1507–1522

Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K et al (2018) AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. CoRR, abs/1810.01943. Retrieved from https://arxiv.org/abs/1810.01943. Accessed 10 Dec 2024

Borg M, Bronson J, Christensson L, Olsson F, Lennartsson O, Sonnsj¨o E et al (2021) Exploring the assessment list for trustworthy ai in the context of advanced driver-assistance systems. 2021 IEEE/ACM 2nd international workshop on ethics in software engineering research and practice (SEthics), Virtual Conference, pp 5–12

Canca C (2019) A new model for AI ethics in R&D. Retrieved from https://www.forbes.com/sites/insights-intelai/2019/03/27/rethinking-ethicsin-ai-rd/. Accessed 20 Jan 2025

Cârlan C, Gauerhof L, Gallina B, Burton S (2022) Automating safety argument change impact analysis for machine learning components. 2022 IEEE 27th Pacific Rim international symposium on dependable computing (PRDC),Beijing, China, pp 43–53

Chatila R, Havens JC (2019) The ieee global initiative on ethics of autonomous and intelligent systems. In: Robotics and well-being. Springer International Publishing, Cham, pp 11–16

Chelouati M, Boussif A, Beugin J, El Koursi E-M (2023) Graphical safety assurance case using goal structuring notation (gsn)—challenges, opportunities and a framework for autonomous trains. Reliab Eng Syst Saf 230:108933

Dawson D, Schleiger E, Horton J, McLaughlin J, Robinson C, Quezada G et al (2019) Artificial intelligence: Australia's ethics framework-a discussion paper. Data61 CSIRO, Australia. https://www.csiro.au/en/research/technology-space/ai/ai-ethics-framework

Digital NSW (2024) Mandatory ethical principles for the use of AI. https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-ethics-policy/mandatory-ethical-principles. Accessed 20 Jan 2025

Dongmo C, Van der Poll JA (2023) An improved user requirements notation (urn) models' construction approach. Systems 11(6):301

Duan L, Rayadurgam S, Heimdahl MP, Ayoub A, Sokolsky O, Lee I (2017) Reasoning about confidence and uncertainty in assurance cases: a survey. In International Symposium on Foundations of Health Informatics Engineering and Systems. Washington, DC, USA. Springer International Publishing, Cham, pp 64–80

Gardner A, Smith AL, Steventon A, Coughlan E, Oldfield M (2022) Ethical funding for trustworthy AI: proposals to address the responsibilities of funders to ensure that projects adhere to trustworthy ai practice. AI Ethics. https://doi.org/10.1007/s43681-021-00069-w

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, Crawford K (2021) Datasheets for datasets. Commun ACM 64(12):86–92

Grigorova S, Maibaum T (2013) Taking a page from the law books: considering evidence weight in evaluating assurance case confidence. In 2013 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Pasadena, CA, USA, pp 387–390

Gusdorf ML (2008) Recruitment and selection: hiring the right person. Society for Human Resource Management, USA, pp 1–14

Han SH, Choi HJ (2022) Checklist for validating trustworthy AI. In 2022 IEEE International Conference on Big Data and Smart Computing (BigComp), Bangkok, Thailand, pp 391–394

Horkoff J, Aydemir FB, Cardoso E, Li T, Mat´e A, Paja E et al (2019) Goal-oriented requirements engineering: an extended systematic mapping study. Requir Eng 24:133–160

IEEE (2022) ISO/IEC/IEEE international standard - systems and software engineering–systems and software assurance–part 2: Assurance case. https://standards.ieee.org/ieee/15026-2/10236/. Accessed 10 Dec 2024

IEEE Computer Society/Artificial Intelligence Standards Committee (C/AISC) (2019) Ieee draft standard for responsible ai licensing (Vol. IEEE P2840). IEEE, New York, USA

IEEE Computer Society/Artificial Intelligence Standards Committee (C/AISC) (2020) Recommended practice for organizational governance of artificial intelligence (Vol. IEEE P2863). IEEE, New York, USA

ITU-T Z.151, I.-T.R. (2018) Z.151 user requirements notation (URN)–language definition. https://www.itu.int/rec/T-REC-Z.151-201810-I/en

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399

Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN et al (2021) Advances and open problems in federated learning. Foundations and Trends® in Machine Learning 14(1–2):1–210

Kaushal N, Kaurav RPS, Sivathanu B, Kaushik N (2023) Artificial intelligence and HRM: identifying future research agenda using systematic literature review and bibliometric analysis. Manag Rev Q 73(2):455–493

Leveson NG (2016) Engineering a safer world: systems thinking applied to safety. The MIT Press, Cambridge, MA, USA

Liao QV, Pribi´c M, Han J, Miller S, Sow D (2021) Question-driven design process for explainable AI user experiences. Preprint at https://arxiv.org/abs/2104.03483. Accessed 10 Dec 2024

Maksimov M, Kokaly S, Chechik M (2019) A survey of tool-supported assurance case assessment techniques. ACM Comput Surv (CSUR) 52(5):1–34

Mansourov N, Campara D (2010) System assurance: beyond detecting vulnerabilities. Morgan Kaufmann, MA, USA

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B et al (2019) Model cards for model reporting. Proceedings of the conference on fairness, accountability, and transparency (FAccT 2019), Atlanta, USA, pp 220–229

Mohamad M, Steghofer J-P, Scandariato R (2021) Security assurance cases—state of the art of an emerging approach. Empir Software Eng 26(4):70

Morley J, Floridi L, Kinsey L, Elhalal A (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26(4):2141–2168

National Artificial Intelligence Centre (2024)Voluntary ai safety standard — guiding safe and responsible use of artificial intelligence in Australia. Australian Department of Industry, Science and Resources, Canberra, Australia

Picardi C, Paterson C, Hawkins RD, Calinescu R, Habli I (2020) Assurance argument patterns and processes for machine learning in safety-related systems. Proceedings of the workshop on artificial intelligence safety (SafeAI 2020), New York, USA, pp 23–30

Radclyffe C, Ribeiro M, Wortham RH (2023) The assessment list for trustworthy artificial intelligence: a review and recommendations. Front Artif Intell 6:1020592

Rinehart DJ, Knight JC, Rowanhill J (2015) Current practices in constructing and evaluating assurance cases with applications to aviation (Tech. Rep.) NASA. https://ntrs.nasa.gov/search.jsp?R=20150002819

Roopalatha N, Sucharita K (2024) Artificial intelligence on human resource management-innovation, challenges and path forward. Educ Adm Theory Pract 30(5):13686–13698

Schwartz R, Schwartz R, Vassilev A, Greene K, Perine L, Burt A, Hall P (2022) Towards a standard for identifying and managing bias in artificial intelligence, vol 3. US Department of Commerce, National Institute of Standards and Technology, USA

Shahandashti KK, Belle AB, Lethbridge TC, Odu O, Sivakumar M (2023) A PRISMA-driven systematic mapping study on system assurance weakeners. Preprint at https://arxiv.org/abs/2311.08328

Sharifi S, Amyot D, Mylopoulos J, McLaughlin P, Feodoroff R (2022) Towards improved certification of complex fintech systems–a requirements-based approach. IEEE 30th international requirements engineering conference workshops (REW 2022), Virtual Conference, pp 205–214

Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S et al (2020) The role of artificial intelligence in achieving the sustainable development goals. Nat Commun 11(1):1–10

Weiss M, Amyot D (2007) Business model design and evolution. In: Challenges in the management of new technologies. World Scientific, Singapore, pp 183–194

Yap RH (2021) Towards certifying trustworthy machine learning systems. Trustworthy AI-integrating learning, optimization and reasoning: First international workshop, tailor 2020, virtual event, september 4–5, 2020, revised selected papers 1 pp 77–82

Zhou J, Chen F (2023) AI ethics: from principles to practice. AI Soc 38(6):2693–2703

Zhou J, Chen F, Berry A, Reed M, Zhang S, Savage S (2020) A survey on ethical principles of ai and implementations. 2020 IEEE symposium series on computational intelligence (SSCI 2020), Canberra, Australia, pp 3010–3017

Zhou J, Gandomi AH, Chen F, Holzinger A (2021) Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics 10(5):593

Zicari RV, Amann J, Bruneault F, Coffee M, Du¨dder B, Hickman E et al (2022) How to assess trustworthy ai in practice. Preprint at https://arxiv.org/abs/2206.09887

Zicari RV, Brodersen J, Brusseau J, Du¨dder B, Eichhorn T, Ivanov T et al (2021) Z-inspection®: a process to assess trustworthy ai. IEEE Trans Technol Soc 2(2):83–97

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.