CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

| | |
|---|---|
| Title | Towards a SAFETY-AI framework for healthcare education |
| Authors | Salim, Kinza;Nana, Vanita Kouomogne;Marshall, Mark T.;Nguyen, Harry |
| Publication date | 2025 |
| Original Citation | Salim, K, Nana, V K, Marshall, M T & Nguyen, H 2025, 'Towards a SAFETY-AI framework for healthcare education', Paper presented at ACML 2025 , Taipei, Taiwan, Province of China, 9/12/25 - 12/12/25. |
| Type of publication | Conference item |
| Rights | © 2025, the Authors. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission. - https://creativecommons.org/licenses/by/4.0/ |
| Download date | 2026-11-03 10:31:41 |
| Item downloaded from | https://hdl.handle.net/10468/19476 |



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Proceedings of Machine Learning Research , 2025

# Towards a SAFETY-AI framework for Healthcare Education

**Kinza Salim** K.SALIM@CS.UCC.IE
**Vanita Kouomogne Nana** NANAKOUOMOGNE.VANITA@UL.IE
**Mark T. Marshall** MARK.MARSHALL@UL.IE
**Hoang D. Nguyen** HN@CS.UCC.IE
*University College Cork, University of Limerick*

## Abstract

Safety is an integral part of healthcare professionalism, and with new technological developments, such as Artificial Intelligence (AI), there is an ongoing need to develop guardrails for healthcare education. The landscape of AI safety frameworks for healthcare education is evolving, with significant development in regulatory compliance, ethical governance, and practical implementation approaches. This paper addresses the need for building a SAFETY-AI framework for healthcare education and proposes a solution towards it. It also provides subjective insights regarding trustworthiness, reliability and the existing concepts of safety in healthcare setups. This work stands as a roadmap for safety in AI practices for healthcare policy makers, educators and clinicians.

**Keywords:** Artificial intelligence; Safety; Education; Healthcare; Reliable AI.

## 1. Introduction

Besides studying various systemic diseases, one of healthcare practitioners' primary focuses is to improve their clinical skills throughout their careers. Human error is a major challenge in healthcare practice. According to WHO statistics from 2023, one in every ten patients is affected by unsafe medical practice (Slawomirski and Klazinga, 2022). There are multiple causative factors which lead to erroneous practice, including human factors such as ineffective clinical handover, delayed team communication, interrupted emergency/crisis identification, and escalation (Delaney et al., 2025).

Another contributory factor that may affect standardized medical practice is misinterpreted technology. In other words, inadequate knowledge of integrated systems or problems faced when using current technical tools can lead to negative outcomes (Holmgren et al., 2023). An example of one such technological factor is AI integration. Although the emergence of AI has reshaped healthcare delivery in areas such as decision-making, diagnosis and workflow management, its adoption in healthcare practices poses a potential threat to ethical integrity and generates real concerns related to privacy, bias, hallucinations and trust (Nana and Marshall, 2025).

One possible way to mitigate these risks is by building a safety net from the very beginning, that is, starting at the point of medical education. To address this, this paper proposes an AI safety framework for healthcare education intended to serve as a pioneering regulatory framework supporting and advancing safe, ethical, reliable and trustworthy AI.

© 2025 K. Salim, V.K. Nana, M.T. Marshall & H.D. Nguyen.

## 2. Related Work

The European Union's AI Act represent a comprehensive regulatory architecture supporting AI safety in healthcare education. While considering healthcare and education as high-risk AI application areas, this legislation advocates for safe, transparent and ethically-compliant AI with a special focus on human rights. Moreover, it mandates AI literacy training for all personnel interacting with AI systems and establishes requirements such as risk assessment, human oversight, and transparency measures. This further reinforces the view that healthcare education requires careful oversight, as a misunderstood or erroneous output from an AI-driven assessment system could contaminate students' knowledge, thereby jeopardizing their academic and professional development (Nana and Marshall, 2025).

Guidance from the World Health Organization (2024) emphasizes the need for human-centered AI development with a particular focus on equity, safety, and transparency in educational applications. The framework highlights the potential benefits of AI systems, including their easy adaptation to user needs, help with building non-technical skills like communication and presentation skills, with a chain-of-thought reasoning that is capability of responding to questions in an efficient manner.

The National Academy of Medicine (NAM) has developed an AI Code of Conduct (AICC) framework to foster the responsible and ethical use of artificial intelligence across healthcare and biomedical research. It is designed as a benchmark framework to help organizations evaluate and align their internal AI governance practices. The framework is built on the Learning Health System (LHS) Shared Commitments, extending these values to the emerging domain of AI. It comprises two central components: The Code Principles, articulating overarching ethical and operational values, and the Code Commitments, translating these principles into actionable guidance for decision-making throughout the AI life cycle. The framework was developed from a comprehensive landscape review of reliable AI literature, existing frameworks from academia, industry, government, and interviews with subject-matter experts to identify commonalities, gaps, and areas of improvement. Grounded in complex adaptive systems theory, the AICC emphasizes that a concise set of shared principles and commitments can drive a coordinated system-level change across the health AI ecosystem (Denecke and Baudoin, 2022; Series et al., 2025). This structure (the AICC) also aims to support trustworthiness and accountability while ensuring that AI applications in healthcare enhance rather than undermine human well-being (Series et al., 2025).

Furthermore, Fjeld et al. (2020) discussed how pre- and post-deployment considerations are important when AI systems are incorporated in public healthcare. They argued that responsible AI principles include assuring ethical, transparent, and accountable usage of technology as per user expectations, corporate values, societal laws and conventions based on responsible AI consensus. Their work shows that a responsive technology should encompass a range of standards that must be satisfied for it to be used. It also projects views regarding the drawbacks of AI systems, such as algorithmic bias, equity concerns, and public health impact.

AI systems are gradually making their way towards medical education as Technology-Enhanced Learning (TEL) frameworks leverage technology for learning purposes. Cook and Ellaway (2015) put special emphasis on efficiency and effectiveness, while also describing how

digital technologies can be used to support educational activities. At the higher education level, new regulations are under development by the Higher Education Authority (HEA) in Ireland regarding the use of AI in higher education settings, where one of their goals is to build an inclusive environment with high-quality public education (O'Sullivan, 2025).

## 3. Definition of Safety in Healthcare

The term "safety" is defined by the WHO as the prevention of errors and adverse effects associated with patients in healthcare (Walton et al., 2010). Adverse events or incidents are part of everyday life, and it is crucial that they are dealt with, especially in a healthcare setting. Safety is generally defined as avoiding doing things that cause harm (commissions) and doing things that prevent harm (omissions). In other words, safety within healthcare entails avoiding harmful actions toward patients and ensuring that necessary actions take place for an appropriate intervention and are not omitted or neglected.

Hollnagel et al. (2015) discuss the concepts of Safety-I and Safety-II. Diving deep into the concept, Safety-I mean factors which led to errors. It is also defined as the absence of adverse outcomes and focuses on identifying and eliminating the root cause of failure. It is reactive in nature and is aimed at preventing things from going wrong. On the other hand, Safety-II is the ability to ensure that as many things as possible go right or have acceptable outcomes (like daily activities). It is a proactive approach which emphasizes the healthcare system's resilience and anticipates adapting to changing conditions.

## 4. Safety in Healthcare Education: Patient and Student Perspectives

When discussing safety in healthcare education, we are referring to safe and ethical learning practices within the subjective learning environment. From the *student's perspective*, safety comprises psychological, pedagogical, ethical and physical factors that ensure that an individual is being protected from any harmful (mis- or dis-) information. Additionally, it also means a student is prepared to receive adequate knowledge necessary to deliver high-quality care, also known as best clinical practice (Perleth et al., 2001).

In practical terms, a safe learning environment is essential. It can enhance students' engagement and trust towards the educational system. Similarly, pedagogical safety will ensure fairness and transparency, therefore reinforcing reliability and institutional trust. Physical and ethical safety will guarantee a logistically suitable and respectful, non-discriminatory learning environment for all healthcare students, thereby strengthening trust between learners and the educational system. Furthermore, a safe learning environment contributes to optimal student performance and is characterized as being free from negative judgments, treating mistakes as learning opportunities, free will to attend to tasks (assisted by no additional pressures like time limitation) and immediate feedback provision to improve best practice (Young et al., 2016).

To develop a trusted healthcare system, consideration of the *patient's perspective* is of utmost importance. It not only promotes patient-centred care but also plays a significant role in high-quality management in a hospital setting. It is highly dependent on communication skills (Hovey et al., 2010) and reliable information delivered in a way that is easily understood by patients themselves and their carers, family members and loved ones.

At the institutional level, a prime goal of clinical educators is to focus on the aspect of patient safety (Aldardeir et al., 2025). A medical curriculum teaches ethical principles and guidelines to address patient concerns and to provide reassurance about existing disease conditions. The American College of Graduate Medical Education and the American Board of Medical Specialities have identified multiple factors which are mandatory for professional competence, which include medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, patient care and systems-based practice (Wu and Busch, 2019). These skills, both technical and non-technical, enhance clinical competencies and lead to an effective, patient-friendly healthcare system.

## 5. What is Reliability, Safety and Trustworthiness in Artificial Intelligence

AI is not a single entity but a spectrum of multi-level intelligent systems and algorithms working together to execute a task, which is evolving in nature and changing with time. Here, we intend to address some frequently interchangeable AI terminologies in the healthcare context.

**Reliable AI** is a term used alongside responsible AI. It is defined as AI that exhibits fairness in ethical groups, is transparent and explainable, and does not misuse confidential datasets. Reliability of an AI system means that it performs an intended task within specified limitations and without any failure. It also means producing the same outputs for the same inputs in a consistent manner (Kaur et al., 2022).

**Safe AI** means no disinformation or hallucination is provided by an AI system. These systems should be robust and resilient in performing tasks as directed by their users. If something goes wrong, a safe AI system is capable of recovering from failure without causing harm to anybody. Moreover, these systems must be capable of dealing with errors at every step of the processing, i.e from prompt generation to task execution (Kaur et al., 2022).

**Trustworthy AI** attributes to protection from adversarial attacks (Chen et al., 2024). EU guidelines for trustworthy AI include respecting all applicable laws and regulations, incorporating ethical principles and values, being robust from a technical perspective, while also taking into account its social environment (Cannarsa, 2021).

## 6. The Need for a Safety-AI Framework in Healthcare Education

Current frameworks often lack validation methodologies, and many educational institutions are still struggling with preliminary issues such as dense curricula, faculty readiness barriers, and limited access to AI tools, particularly at under-resourced institutes (Saroha, 2025). Considering these issues, there is a need to develop and deploy a robust, reliable, trustworthy and readily available AI framework tailored to support safe AI practices at the institutional level, while accounting for the dynamic needs of multidisciplinary teams, which include students, healthcare educators, AI researchers and policy makers. This enables seamless collaboration among different stakeholders, allowing data-driven decision-making and innovation in healthcare education.

From an educational and healthcare viewpoint, any such framework must be grounded in principles of safe AI ensuring transparency, fairness, integrity and ethical compliance in all interactions. By promoting responsible AI use, the framework can enable students not only to benefit from advanced learning tools but also to gain critical awareness of the societal and professional implications of AI in healthcare and research. An initiative was taken in this direction by the European Commission's High-Level Expert Group on AI (HLEG-AI), which created an assessment list for Trustworthy Artificial Intelligence (ALTAI) tools that included seven key requirements. These were: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability (Radclyffe et al., 2023). A framework based on these factors can both be standardized and promote safe healthcare practice.

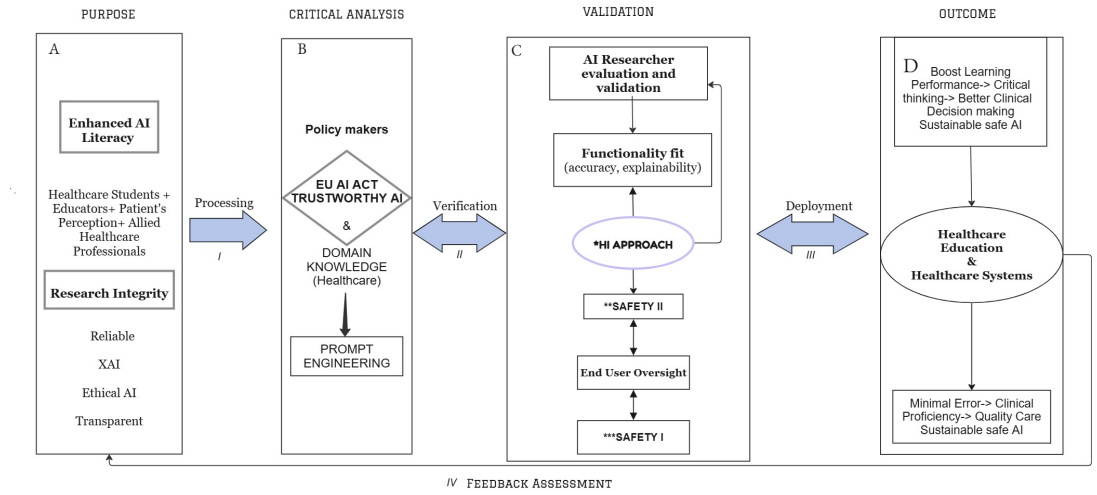



Figure 1: SAFETY-AI Framework for Healthcare Education. **HI: Hybrid Intelligence (Human in the loop approach), **Safety II: commissions/cause harm ***Safety I:omissions/prevent harm.*

### 6.1. The SAFETY-AI Framework

The usefulness of a framework lies in its effectiveness, applicability, functional fit and outcome relevance. Our proposed framework distinguishes itself through its comprehensive conceptual architecture. It introduces a multidisciplinary design that bridges educational, clinical, technical and policy domains, enabling seamless collaboration among students, educators, and AI researchers. The granularity of its architecture enables easy understanding and interpretability.

Our framework integrates adaptive learning technologies with a feedback mechanism which supports personalized learning and decision-making. Moreover, its alignment with the EU AI Act reinforces its ethical ideology and safety-AI principles, ensuring the integration

of safe AI practices. Additionally, it also redefines how diverse stakeholders engage with AI systems at the institutional, healthcare and community levels.

The architecture of our framework is illustrated in Figure 1, focusing on the conceptual aspect of the framework by illustrating the most relevant components. The design and interactions among stakeholders are illustrated at a granular level. Our work integrates the concept of Human-in-the-Loop (HITL) as a central idea. This ensures that AI systems are assistive and robust but requires a human's critical analysis before deployment.

Modern educational needs have emphasized AI as an integral part of the curriculum (Pedro et al., 2019). Our conceptual SAFETY-AI framework encompasses the requirements of students and educators for safe healthcare practices that serve as a necessity of the modern era.

### 6.2. Components of the Framework

The SAFETY-AI framework has four basic pillars denoted as Purpose, Critical Analysis, Validation and Outcome and four stages of transitions: Processing, Verification, Deployment, and Feedback Assessment, which serve efficiently at both teaching and learning prospects of healthcare. As observed in the literature, AI has gained momentum in healthcare education with applications in medical (Masters, 2019), nursing (Simms, 2025) and allied medical professions (Hoffman et al., 2024).

#### 6.2.1. Purpose

The proposed framework is intended to enhance AI literacy within healthcare education. There is a need to understand and expand knowledge of AI systems as this will help foster innovation and strengthen research. For students, patients and healthcare educators, the deployment of this framework will enhance research integrity, as well as enabling reliable, explainable, ethical, and transparent AI in healthcare practice, therefore reinforcing safety principles.

One challenge of AI integration in healthcare is uncertainty. It is non-trivial for the user to predict when the system will be inaccurate, even more so to perceive the inaccuracy. As a result, it is crucial to clearly redefine how users interact with AI within healthcare, and healthcare education specifically. There is a need for clarification regarding the intended use of AI systems, their benefits, their limitations and their potential impact in clinical decision making.

While envisioning safe AI practices within healthcare education, this framework enables stakeholders (healthcare professionals) to harness the benefits of AI in healthcare education. Nevertheless, it highlights the need to raise awareness about the risks to patient safety and the importance of the EU AI Act in order to build trust and transparency in healthcare technological adaptations. It also enhances safety AI principles and links them directly to allied healthcare professionals like nurses, healthcare assistants, occupational therapists, healthcare administrators, physiotherapist, etc., leading to generalisability across the healthcare system.

#### 6.2.2. Critical Analysis

This component of the framework is centered on requirements for safe AI practice within healthcare education. Its main pillar is the EU AI ACT, but it also includes important elements such as healthcare domain knowledge, policymakers and prompt engineering. The critical analysis component is underpinned by the EU AI ACT Edwards (2021), ensuring AI systems within healthcare are safe and trustworthy. To promote trustworthy AI practices within healthcare, the EU AI Act puts emphasis on the development, deployment and utilisation of AI medical systems. It classifies risks associated with the use of AI systems into four categories: unacceptable, high, limited and minimal. It also defines healthcare compliance obligations, including but not limited to transparency, data governance, risk management, AI system technical documentation and human oversight (Busch et al., 2024).

Acting on the EU regulation for trustworthy AI in healthcare is required and recommended. However, within healthcare education, stakeholders (mainly students, patients, and educators) ought to take into consideration essential aspects like healthcare domain knowledge and prompt engineering. Interacting with high-risk AI healthcare systems, such as emergency triage tools or Large Language Models (LLMs) generating diagnosis reports and recommending treatment plans, requires a certain level of expertise and advanced knowledge in both the medical and AI domains. The rapid rise of generative AI in healthcare has accelerated the adoption of LLMs, particularly in educational settings (Nana and Marshall, 2025). In such systems, prompt engineering is the process of refining instructions or inputs given to an LLM in order to obtain a desired output. This would appear to be a skill required to optimally use Generative AI systems in sensitive areas like healthcare education. Enhancing safety AI in the healthcare educational context requires an ensemble of elements as described above; however, policymakers also have an important role to play at this stage. Beyond the EU AI ACT, they have to remain updated with relevant medical and technical knowledge or expertise, as this can be a pivotal advantage in the development and deployment of trustworthy AI principles and practices in healthcare.

#### 6.2.3. Validation

Safety in AI is a joint responsibility of both the researchers/developers and the end users. Our framework is built on the Hybrid Intelligence ideology of combining both human intelligence and AI when interacting with AI systems. The validation component of our framework encompasses a prior evaluation from both the AI researchers and the end users before validation against established regulations. This essentially implies checking and validating the AI-generated information before considering it for healthcare decision-making.

For safe AI practices, our framework recommends validation as a bi-directional approach. This means that validation is the responsibility of both the AI developer and also of the end user. From the developer's perspective, the AI system should be technically assessed to ensure fault tolerance and reliability (functionality fit) and zero downtime performance.

From an end-user perspective, our framework integrates validation to the levels of both Safety II (factors that cause harm) and Safety I (factors that prevent harm), denoting the end user's responsibility in ensuring that any AI system output is medically sound and compliant with medical ethics and the EU AI Act. Within an educational setting, this would mean validating the output before dissemination to students.

#### 6.2.4. Outcome

The Outcome pillar includes a range of areas in which our framework can be of benefit within healthcare education.

**Building AI literacy.** Our framework plays a pivotal role in improving AI literacy by equipping individuals with the knowledge and skills necessary to understand, evaluate, and use AI-generated data responsibly. AI literacy is increasingly recognized as a foundational skill in healthcare education and in professional development. It enables users to critically engage with AI technologies. There are three existing theories that guided the development of this framework: Self-Paced Learning (SPL), the Technology Acceptance Model (TAM), and the Theory of Planned Behaviour (TPB). These theories pave the way for the SAFETY-AI framework. Self-Paced Learning (SPL) connects self-regulation with planning, monitoring, and evaluating academic activities. It directly ties to how students manage their adherence to institutional policies and ethical standards (Samonte et al., 2025).

The Technology Acceptance Model (TAM), which was introduced by Fred Davis in 1989, explores the adoption of new technology and new services by using electronics; this pattern proposes acceptance of new technologies in the modern era (Davis, 1989).

The Theory of Planned Behaviour (TPB) is a key framework to comprehend awareness of institutional policies that can influence students' compliance with academic integrity. According to the TPB, three factors determine an individual's objectives. Attitudes: A student's positive or negative evaluations of adhering to academic integrity. Subjective norms: The perceived societal pressures to adhere to institutional policies and perceived behavioural control: The student's belief in their capacity for adhering to academic integrity ideals, given their awareness of policies (Ajzen, 1991).

These theories consolidate the fact that student perspective is important and to develop trust in AI their perception holds an immense significance. Another important member at academic institute is the mentor or educator who needs to know about AI systems.

**Strengthening the concept of reliable AI and research integrity.** Protecting research integrity is a prime goal of academics. Meanwhile, building trust in AI is the responsibility of AI researchers. This framework unifies both perspectives and also adds an additional layer of safety by including the EU Trustworthy concept of AI at the Analysis phase of processing.

**Impact on Healthcare Education.** Boosting a learning curve means a better understanding level and confidence building. This framework aims to emphasize critical thinking and improve cognitive ability.

**Improved Decision-Making in Healthcare.** For an efficient healthcare system, there is a need to minimize errors. With human oversight at the Analysis and Validation stages, our framework ensures safety in medical practice.

**Multi-disciplinary collaboration.** Collaboration is a cornerstone of the SAFETY-AI framework, ensuring that AI systems are an assistive tool, rather than a replacement

for human judgment. Efficient clinical surveillance and research oversight can provide timely intervention if AI systems face problems, such as data bias or hallucinations.

**Feedback Assessments.** For further refinement of the framework, a feedback assessment method is added to keep a track record of the progression through the stages. A questionnaire about what improvements are needed for the framework would be conducted as a quantitative analysis to evaluate the effectiveness of the framework.

These features show that the adoption of a robust SAFETY-AI framework can maximize efficiency levels in teaching and learning.

## 7. Discussion and Future Roadmap

The current healthcare educational landscape presents significant opportunities for the Safety-AI framework. It capitalizes on identified gaps and proposes approaches that create innovative solutions for safe care and learning. Collaborative efforts with Human-Centered AI and Explainable AI (XAI) can additionally help educators to understand how feedback loops work in AI systems and can demystify problems such as the black box effect (Kaur et al., 2022). The emergence of AI systems has brought transformation and innovation to healthcare in recent years. From an AI researcher's viewpoint, healthcare AI systems were initially intended to support healthcare practitioners in performing multiple tasks, such as diagnostics (Sadr et al., 2025) or AI-assisted robotic surgery (Knudsen et al., 2024), but concerns related to integrity and autonomy have emerged, hindering trust in AI systems. For example, the emergence of LLMs in healthcare education has highlighted limitations, including data privacy, confidentiality, hallucinations, and trust (Zhui et al., 2024). Although consistent efforts are made to reduce distrust in AI (Starke and Ienca, 2024), little success has been achieved. There have been other attempts at proposing frameworks for safe AI use in healthcare. For example, a three-tier approach was proposed that provides clinicians with personalised real-time insights (De Silva et al., 2025). Another framework was designed with stakeholder safety at its core, requiring developers to demonstrate biological, psychological, economic, and social values Khan and Seto (2023). This framework specifically addressed the feature of the Hippocratic Oath, i.e *do no harm* (Sokol, 2013). As the prime focus in patient safety, there are frameworks which are patient-centred in origin, for instance governance model for AI in healthcare (Reddy et al., 2020). It has four basic components, which include fairness, transparency, trustworthiness and accountability. This framework deals with the ideology of data governance, patient involvement and clinicians, but does not include any collaboration with researchers. However, still no approach in particular is held as a standardized framework for healthcare education. Our SAFETY-AI framework is novel as it interlinks cross-functional team involvement, keeps human oversight, and incorporates governance policies for safe AI data transmission. Secondly, it has an evaluation feedback system where data can be recorded for further iteration and improvement. This feature captures the sustainability factor. The framework addresses the individual, organizational and institutional levels and holds the promises a positive impact in society. It also helps in improving AI literacy at for educators and students. The limitations of this research lies in the theoretical nature of the proposed framework. Its deployment requires operational resources that extend beyond theoretical considerations. These include technical resources

such as AI safety specialists and cybersecurity experts; As well as educational resources that support AI literacy training, faculty improvement and the upskilling required for the safe and responsible use AI technologies (e.g.,new curriculum components, faculty time, expertise and educational resources, computational resources). Clear communication of potential risks is also essential. Additionally, the deployment of the proposed framework relies on human resources including domain experts (e.g.,data governance, legal and compliance specialists), ethicists who can strengthen fairness and human oversight mechanisms, and social scientists capable of assessing its broader societal impact. Future work includes the development and implementation of a case study capable of showcasing the framework in action, moreover, assessing its feasibility and effectiveness to facilitate its practical applicability and validation.

## 8. Conclusion

Safety is grounded in the principle of "First, do no harm" (Sokol, 2013), which remains central to healthcare practice. The increasing adoption of AI in healthcare education necessitates a robust safety framework to ensure trustworthiness, which strengthens research integrity and ensures reliability. This paper highlighted the need for a SAFETY-AI Framework to guide responsible AI adoption in medical education and practices. Additionally, our work proposed a solution towards the development and deployment of a SAFETY-AI framework at the healthcare educational level. Future efforts require multidisciplinary collaboration, including AI researchers, policymakers, and educators, all working together to build safe healthcare AI systems. This effort fosters a culture of inclusion along with safety as an integral component and keeps patient-centred care as its prime goal.

## Acknowledgments

This publication has emanated from research supported in part by grants from Research Ireland under Grant [12-RC-2289-P2] and [18/CRT/6223] which is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

Njoud Aldardeir, Qabirul Karan Abdullah, and Linda Jones. Patient safety education in undergraduate medical education through a global lens: a scoping review. *BMC Medical Education*, 25(1):544, 2025.

Felix Busch, Jakob Nikolas Kather, Christian Johner, Marina Moser, Daniel Truhn, Lisa C Adams, and Keno K Bressem. Navigating the european union artificial intelligence act for healthcare. *npj Digital Medicine*, 7(1):210, 2024.

Michel Cannarsa. Ethics guidelines for trustworthy ai. *The Cambridge handbook of lawyering in the digital age*, pages 283–297, 2021.

Chen Chen, Xueluan Gong, Ziyao Liu, Weifeng Jiang, Si Qi Goh, and Kwok-Yan Lam. Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations. *arXiv preprint arXiv:2408.12935*, 2024.

David A Cook and Rachel H Ellaway. Evaluating technology-enhanced learning: a comprehensive framework. *Medical teacher*, 37(10):961–970, 2015.

Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989.

Chameera De Silva, Thilina Halloluwa, and Dhaval Vyas. A multi-layered research framework for human-centered ai: Defining the path to explainability and trust. *arXiv preprint arXiv:2504.13926*, 2025.

Aoife Delaney, Suzanne McCarthy, Kirstyn James, and Aoife Fleming. Using a human factors and ergonomic methodology to improve medication safety outcomes in a hospital setting; a systematic review. *Human Factors in Healthcare*, page 100112, 2025.

Kerstin Denecke and Claude R Baudoin. A review of artificial intelligence and robotics in transformed health ecosystems. *Frontiers in medicine*, 9:795957, 2022.

Lilian Edwards. The eu ai act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*, 1:25, 2021.

Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, (2020-1), 2020.

Jane Hoffman, Laetitia Hattingh, Lucy Shinners, Rebecca L Angus, Brent Richards, Ian Hughes, and Rachel Wenke. Allied health professionals' perceptions of artificial intelligence in the clinical setting: Cross-sectional survey. *JMIR formative research*, 8:e57204, 2024.

Erik Hollnagel, Robert L Wears, Jeffrey Braithwaite, et al. From safety-i to safety-ii: a white paper. *The resilient health care net: published simultaneously by the University of Southern Denmark, University of Florida, USA, and Macquarie University, Australia*, 2015.

AJ Holmgren, S McBride, B Gale, and S Mossburg. Technology as a tool for improving patient safety. *Technology as a Tool for Improving Patient Safety*, 2023.

RB Hovey, A Morck, S Nettleton, S Robin, D Bullis, A Findlay, and H Massfeller. Partners in our care: patient safety from a patient perspective. *Quality and Safety in Health Care*, 19(6):e59–e59, 2010.

Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2):1–38, 2022.

Waqas Ullah Khan and Emily Seto. A "do no harm" novel safety checklist and research approach to determine whether to launch an artificial intelligence–based medical technology: Introducing the biological-psychological, economic, and social (bpes) framework. *Journal of medical Internet research*, 25:e43386, 2023.

J Everett Knudsen, Umar Ghaffar, Runzhuo Ma, and Andrew J Hung. Clinical applications of artificial intelligence in robotic surgery. *Journal of robotic surgery*, 18(1):102, 2024.

Ken Masters. Artificial intelligence in medical education. *Medical teacher*, 41(9):976–980, 2019.

Vanita Kouomogne Nana and Mark T Marshall. Generative artificial intelligence in healthcare: A bibliometric analysis and review of potential applications and challenges. *AI*, 6 (11):278, 2025.

James O'Sullivan. Generative ai in higher education teaching and learning. 2025.

Francesc Pedro, Miguel Subosa, Axel Rivas, and Paula Valverde. Artificial intelligence in education: Challenges and opportunities for sustainable development. 2019.

Matthias Perleth, Elke Jakubowski, and Reinhard Busse. What is 'best practice'in health care? state of the art and perspectives in improving the effectiveness and efficiency of the european health care systems. *Health policy*, 56(3):235–250, 2001.

Charles Radclyffe, Mafalda Ribeiro, and Robert H Wortham. The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in artificial intelligence*, 6:1020592, 2023.

Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. A governance model for the application of ai in health care. *Journal of the American medical informatics association*, 27(3):491–497, 2020.

Hossein Sadr, Mojdeh Nazari, Zeinab Khodaverdian, Ramyar Farzan, Shahrokh Yousefzadeh-Chabok, Mohammad Taghi Ashoobi, Hossein Hemmati, Amirreza Hendi, Ali Ashraf, Mir Mohsen Pedram, et al. Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. *European Journal of Medical Research*, 30(1):418, 2025.

Blissa Mae V Samonte, Annabelle P Acedera, et al. Artificial intelligence literacy, awareness of institutional policies and self-regulation skills as predictors of students'adherence to academic integrity. *International Journal Of All Research Writings*, 6(11):245–254, 2025.

Sarup Saroha. Artificial intelligence in medical education: Promise, pitfalls, and practical pathways. *Advances in Medical Education and Practice*, pages 1039–1046, 2025.

The Learning Health System Series, Sunita Krishnan, Michael Matheny, Elaine Fontaine, and Laura Adams. Ai code of conduct framework. In *An Artificial Intelligence Code of Conduct for Health and Medicine: Essential Guidance for Aligned Action*. National Academies Press (US), 2025.

Rachel C Simms. Generative artificial intelligence (ai) literacy in nursing education: A crucial call to action. *Nurse Education Today*, 146:106544, 2025.

Luke Slawomirski and Niek Klazinga. The economics of patient safety: from analysis to action. 2022.

Daniel K Sokol. "first do no harm" revisited. *Bmj*, 347, 2013.

Georg Starke and Marcello Ienca. Misplaced trust and distrust: how not to engage with medical artificial intelligence. *Cambridge Quarterly of Healthcare Ethics*, 33(3):360–369, 2024.

Merrilyn Walton, Helen Woodward, Samantha Van Staalduinen, Claire Lemer, Felix Greaves, Douglas Noble, Benjamin Ellis, L Donaldson, Bruce Barraclough, et al. The who patient safety curriculum guide for medical schools. *BMJ Quality & Safety*, 19(6): 542–546, 2010.

World Health Organization. *Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance*. World Health Organization, 2024.

Albert W Wu and Isolde M Busch. Patient safety: a new basic science for professional education. *GMS journal for medical education*, 36(2):Doc21, 2019.

JE Young, MI Williamson, and TG Egan. Students' reflections on the relationships between safe learning environments, learning challenge and positive experiences of learning in a simulated gp clinic. *Advances in Health Sciences Education*, 21(1):63–77, 2016.

Li Zhui, Li Fenghe, Wang Xuehu, Fu Qining, and Ren Wei. Ethical considerations and fundamental principles of large language models in medical education. *Journal of Medical Internet Research*, 26:e60083, 2024.